ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN INGENIERÍA INFORMÁTICA

# CLASIFICADOR SEMÁNTICO DE MEDIOS SOCIALES MEDIANTE ANÁLISIS BASADO EN EXPRESIONES: APLICACIÓN A BLOGS Y DIARIOS

# SEMANTIC CLASSIFIER OF SOCIAL MEDIA BY ANALYSIS BASED ON EXPRESSIONS: APPLICATION TO BLOGS AND JOURNALS

Realizado por
**Ramón Guevara Quesada**
Tutorizado por
**Dr. José Ignacio Peláez Sánchez**
Departamento
**Lenguajes y Ciencias de la Computación**

UNIVERSIDAD DE MÁLAGA
MÁLAGA, (diciembre 2014)

Fecha defensa:
El Secretario del Tribunal

**Resumen y palabras claves**

La reputación corporativa es la percepción que los diferentes públicos de una empresa tienen acerca de la misma. Estos públicos o stakeholders, son aquellos que se relacionan con la empresa de cualquier manera, pudiéndose distinguir entre públicos internos, empleados, accionistas, directivos, …; y públicos externos, clientes, proveedores, fondos de inversión, ….

En los últimos años estamos presenciando como las empresas e instituciones dedican importantes esfuerzos a desarrollar políticas de Reputación Corporativa, ya que una disminución de la reputación redunda directamente en la cuenta de resultados de las compañías por lo que las mejoras que se introduzcan en la gestión de la reputación redundaran directamente en su sostenibilidad económica y social del país donde operan.

Para poder medir la RC de una empresa o institución es preciso determinar el sentimiento que dicha empresa o institución tiene en sus públicos. Pero llevar a cabo esta medición es cada día más difícil, ya que la sociedad de la información donde estamos inmersos, hace que los canales de comunicación de los públicos sean mucho mayores que hace unos años( redes sociales, blog, diarios, foros, …) y además, la comunicación puede hacerse casi en tiempo real, tal como sucede en las redes sociales.

Para hacer estas mediciones, las principales empresas del sector han utilizado principalmente encuestas las cuales las realizaban empresas especializadas, lo que supone un alto coste y además, los resultados son obtenidos a tiempo pasado o no se alcanza un tamaño de población significativa.

El objetivo de este trabajo es realizar una aplicación para determinar la reputación de una empresa en medios de comunicación online. Para ello, se ha desarrollado un sistema de lectura de medios online, que permite localizar y extraer la información de los medios de comunicación online; un clasificador semántico para analizar la información recogida y clasificarla en diferentes temáticas extrayendo el sentimiento de los textos; y finalmente, una interfaz para interactuar con el usuario.

**Palaras claves:**
Reputación corporativa, RC, análisis semántico, clasificador semántico, diario, blog, stakeholders, Alva in Depth, RepTrak®, TRI * M, Eclipse, JBoss Tools, Hibernate, ROME, Netbeans, Glassfish, SQL Server 2012, social media, web 2.0, sentiment, información, interfaz, RSS, público, empresa, Merco, Expresiones, big data, Java, JSP.

Corporate reputation is the perception different public have of a company. These public are those that relate to the company in any way, they being able to distinguish between internal audiences, employees, shareholders, directors, ...; and external audiences, customers, suppliers, investment funds ....

In recent years we are witnessing how businesses and institutions devote significant efforts to develop policies of corporate reputation, since a decrease of reputation directly affects the bottom line of the companies making the improvements introduced in the management of reputation redound directly to their economic and social sustainability of the country where they operate.

To measure the RC of a company or institution must determine the sentiment that the company or institution has on its publics. But to perform this measurement is becoming increasingly difficult, as the information society in which we operate, makes the communication channels of the public are much greater than a few years ago( social networking, blogs, journals, forums, ...) and also, the communication can be done in near real time, such as in social networks.

To make these measurements, the major companies have used mainly surveys conducted by specialized companies, which implies a high cost and also the results are obtained last time or no significant population size is achieved.

The objective of this work is an application to determine the reputation of a company in online media. To do this, we have developed a system for reading online media, enabling you to locate and extract information from online media; semantic classifier to analyze and classify the information collected in different thematic extracting the text sentiment ; and finally, an interface for user interaction.

**Keywords:**
Corporate Reputation, CR, semantic analysis, SEMANTIC CLASSIFIER, journal, blog, stakeholders, Alva in Depth, RepTrak®, TRI * M, Eclipse, JBoss Tools, Hibernate, ROME, Netbeans, Glassfish, SQL Server 2012, social media, web 2.0, sentiment, information, interface, RSS, public, company, Merco, Expressions, big data, Java JSP.

# Índice de contenido

## Índice de figuras

## 1. Introducción

### 1.1 Motivación del proyecto

La reputación corporativa (RC) es la percepción que tienen los diversos públicos que interactúan con la empresa entre los que se pueden distinguir los accionistas, clientes, empleados y otros grupos de interés. Estas percepciones son el resultado del comportamiento desarrollado por la empresa a lo largo del tiempo y describe su capacidad para cubrir las expectativas y aportar valor a los mencionados grupos.

La pérdida de RC impacta directamente en los resultados de las compañías, por lo que las mejoras que se introduzcan en la gestión de su RC redundan directamente en su sostenibilidad y en la del país donde operan.

Por este motivo, son cada vez más las empresas que se preocupan por generar una buena RC con sus públicos, lo que les está llevando a desarrollar políticas que integran en su desempeño la responsabilidad social, la ética empresarial, el buen gobierno corporativo, las relaciones sostenibles y de confianza, la calidad en los procesos, que impactan en el negocio (cuestiones medioambientales, normativas, del mercado, etc.) y modelos de desarrollo profesional entre otros.

La aparición de la web 2.0 ha permitido a los internautas convertirse en un participante activo en cuanto a la creación de contenido, dando lugar a una sociedad que participa, comunica y genera contenido. Algunas de las fuentes de información (también llamados fuentes sociales) que han surgido son los blogs, las wikis y las redes sociales (social media), prensa digital, etc.

El desarrollo de la Web 2.0 en la red de redes a diferencia de la Web de finales del siglo pasado, destaca en que los individuos pueden tener tanta importancia como las empresas o los medios de comunicación. Este entorno en el que lo importante son las personas, está teniendo una influencia cada día mayor en la sociedad y en la economía tal y como las conocemos.

Según datos de un estudio realizado por el INE sobre la Sociedad de la Información, el 80% de los contenidos existentes en Internet en los años noventa estaba creado por empresas y medio de comunicación y que tan sólo el 20% restante había sido creado por los usuarios. Además, el objetivo de la Web se centraba en ser prácticamente una galería comercial con anuncios, escaparates y tiendas. La participación del usuario final era mínima, algo que ha cambiado mucho.

La incorporación de las TIC (Tecnologías de la Información y la Comunicación) a la gestión empresarial mejora la competitividad y consigue una imagen mucho más sólida e innovadora de cara al exterior. La apuesta consiste en saber adoptar una nueva cultura empresarial, coherente, por supuesto, con la era de Internet y con las nuevas tecnologías que facilitan el posicionamiento en el mercado de distintas empresas.

Los medios sociales son ricos en la influencia y la interacción entre pares y con una audiencia pública que es cada vez más «inteligente» y participativa. El medio social es un conjunto de plataformas digitales que amplía el impacto del boca a boca y también lo hace medible y, por tanto, rentable por medio de la mercadotecnia de

medios sociales y el CRM social.

Los responsables de comunidad se encargan de crear y cuidar las comunidades en torno a las empresas generando contenido de valor, creando conversación, animando a las personas a participar, monitorizando la presencia en la red de las marcas, etc. Los medios sociales han cambiado la comunicación entre las personas, y entre las marcas y las personas.

Esto ha llamado la atención de las empresas, ya que estas fuentes de información se han convertido en un lugar de donde sacar las opiniones que tiene el público hacia su empresa( ya sean sus productos, política, etc). Estas opiniones juegan un papel crucial en la toma de decisiones, ya que cuando adquirimos un producto o servicio lo hacemos influenciados por las opiniones de los demás usuarios y la experiencia que han tenido con estos.

De todo esto nace el concepto de 'clasificador semántico' o 'minería de opinión', que consiste en el uso de la inteligencia artificial para obtener de forma automática información útil acerca de las opiniones, preferencias y tendencias de los usuarios de las redes sociales.

Usando estás técnicas se puede extraer las opiniones de:
- Eventos sociales.
- Movimientos o partidos políticos.
- Estrategias comerciales de empresas.
- Imagen de marca.
- Productos y servicios.
- Personalidades relevantes.
- Etc.

Una vez extraída la información la minería de opinión también permite valorarla cuantitativamente, de manera que podemos saber si se está hablando de forma positiva, negativa o neutra y también ser capaces de medir la intensidad de dicha opinión.

La dificultad en la minería de opinión radica en que las opiniones al estar escritas en lenguaje natural es necesario un gran conocimiento de las reglas semánticas de un idioma, así como de la sintaxis, el léxico y la gran variabilidad existente en los mensajes informales. Esta característica hace que las herramientas existentes no tengan todavía el rendimiento deseado ni puedan aplicarse indistintamente a diferentes idiomas.

## 1.2 Estudio del arte

### 1.2.1 Introducción

El concepto de reputación corporativa hizo su primera aparición en 1983, presentado por la revista Fortune en su ranking de las empresas más admiradas. En paralelo, comenzaron a identificarse los distintos públicos que evalúan el accionar de una empresa y que unos años después darían lugar al concepto de stakeholders. El estudio del tejido social de la empresa, y su creciente complejidad, se convertía en tema de discusión en el ámbito del management.
Stakeholders es un término inglés utilizado por primera vez por Edward Freeman en

su obra: "*Strategic Management: A Stakeholder Approach*" y puede definirse como cualquier persona o entidad que es afectada o concernida por las actividades o la marcha de una organización; por ejemplo, los trabajadores de esa organización, sus accionistas, las asociaciones de vecinos afectadas o ligadas, los sindicatos, las organizaciones civiles y gubernamentales que se encuentren vinculadas, etc.

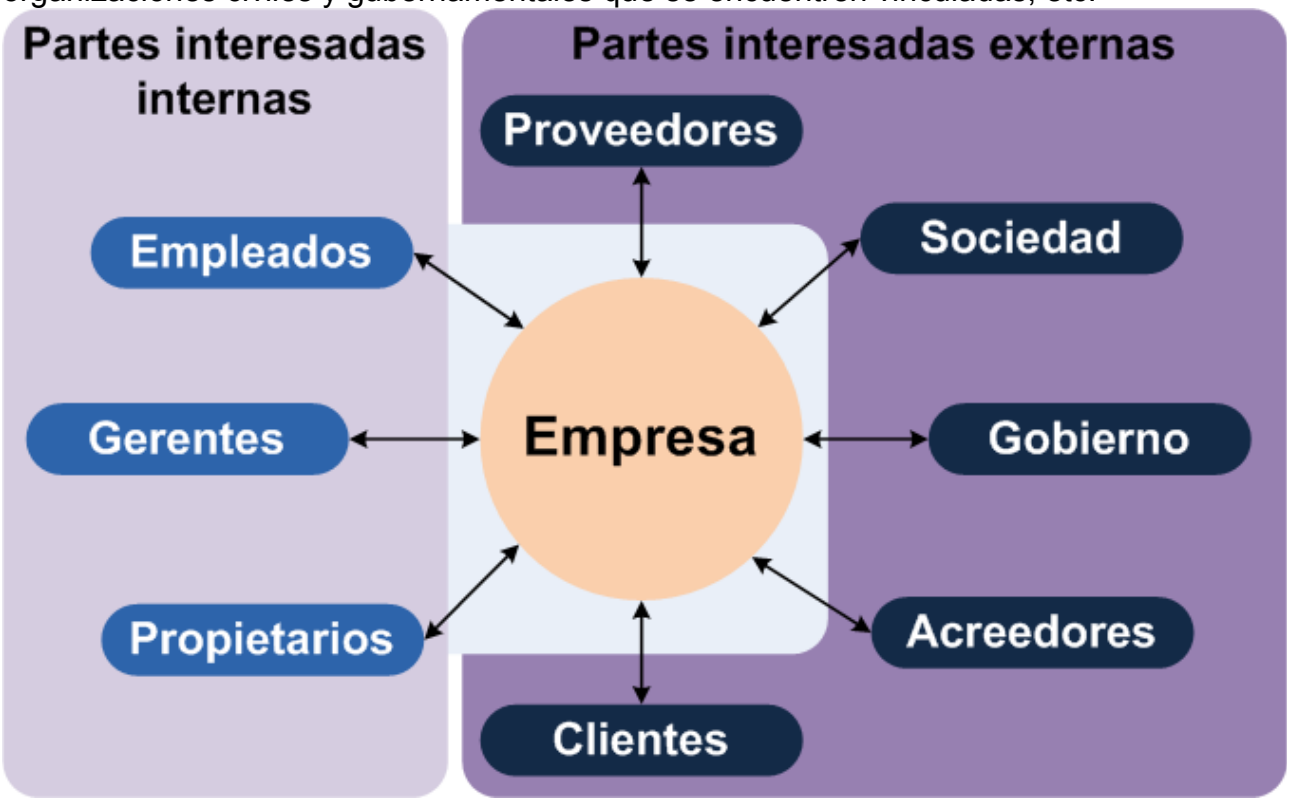



Figura 1. Partes interesadas de una empresa

Hacia los años noventa, y como complemento de la reputación, surge la noción de sustentabilidad corporativa de la mano de las organizaciones ambientales. Es en 1994 cuando el reconocido académico británico John Elkington, introduce el concepto de Triple Bottom Line, para referirse al desempeño de las empresas en tres dimensiones: económica, social y ambiental. Finalmente, en 1997, el programa ambiental de las Naciones Unidas (ONU), junto con distintas organizaciones ambientalistas, crea el Global Reporting Initiative(GRI), con el fin de determinar un estándar mundial en los reportes de sustentabilidad de las empresas y lograr equipararlos a los informes financieros.

Su naturaleza perceptual: La reputación corporativa representa lo que la sociedad siente y piensa de una empresa, basándose en la información o desinformación que tenga de sus actividades; ambiente de trabajo, rentabilidad pasada y proyecciones futuras. Es por su naturaleza perceptual que las empresas no pueden ejercer control directo sobre su reputación. Sin embargo, la reputación corporativa es un activo gestionable. La imagen de la organización en el mercado debe cuidarse y construirse en cada interacción con los stakeholders. Durante años, la empresa British Petroleum(BP) desarrolló acciones para distinguirse como empresa petrolera al cuidado del medio ambiente. Su vinculación con programas ambientales, su compromiso en la reducción de emisiones de gases y sus fuertes inversiones en estudios de energía alternativa, le permitieron desarrollar una imagen positiva en la sociedad. Pero, lamentablemente, el derrame de petróleo ocurrido en 2010 en el

Golfo de México, echó por tierra tantos años de esfuerzo. Sin embargo, no hay dudas de que la reputación construida por BP previa al accidente amortizó el golpe y actuó como reserva de buena voluntad, dándole un margen de confianza mayor frente a sus stakeholders.

Podemos distinguir dos tipos de stakeholders dependiendo de la influencia que tienen sobre la empresa: 'primarios' que tiene una influencia directa y son fundamentales para la existencia de la empresa; y los 'secundarios' que se encuentran en el entorno de la empresa e influyen en los primarios. Cada grupo tiene sus propios asuntos sociales y económicos, pero aún así interactúan entre ellos y el resto de la sociedad transmitiendo su percepción y creando una imagen pública.

Los empleados, como dueños del talento que la empresa necesita, deben ser atraídos y motivados. A su vez, la empresa debe convencer a sus clientes de comprar sus productos entre las múltiples alternativas que el mercado les ofrece. En cuanto a los accionistas, ellos son los que proveen los recursos para financiar las iniciativas de la empresa, si ésta no los convence, la misma supervivencia puede estar en juego. Y, por último, si la comunidad donde opera la empresa no apoya sus proyectos, existe el riesgo de que los obstaculice.

Por tanto, las iniciativas de una empresa son exitosas si aportan credibilidad a su estrategia y logran el consenso de sus stakeholders. En el tiempo, las experiencias que ellos tengan con la empresa, determinarán la calidad de su relación y ésta se irá cristalizando en un activo intangible: la reputación corporativa.

### 1.2.2 Big Data

El primer problema que se nos presenta y del que vamos a hablar a continuación es la gran cantidad de información que tenemos que capturar, analizar, búsqueda y la visualización, el término que hace referencia a esta definición es Big Data (del inglés grandes datos).

Big Data como hemos dicho hace referencia al manejo de grandes volúmenes de datos. En 2012 se dimensionaba su tamaño en una docena de terabytes hasta varios petabytes de datos.

El límite superior de procesamiento se ha ido desplazando a lo largo de los años, de esta forma los límites que estaban fijados en 2008 rondaban los órdenes de petabytes a zettabytes de datos. Los científicos con cierta regularidad encuentran limitaciones debido a la gran cantidad de datos en ciertas áreas, tales como la meteorología, la genómica, la conectómica, las complejas simulaciones de procesos físicos, y las investigaciones relacionadas con los procesos biológicos y ambientales, Las limitaciones también afectan a los motores de búsqueda en internet, a los sistemas finanzas y a la informática de negocios. Los *data sets* crecen en volumen debido en parte a la introducción de información ubicua procedente de los sensores inalámbricos y los dispositivos móviles (por ejemplo las VANETs), del constante crecimiento de los históricos de aplicaciones (por ejemplo de los logs), cámaras (sistemas de teledetección), micrófonos, lectores de radio-frequency identification. La capacidad tecnológica per-cápita a nivel mundial para almacenar datos se dobla aproximadamente cada cuarenta meses desde los años ochenta. Se estima que en 2012 cada día fueron creados cerca de 2,5 trillones de bytes de datos (del inglés *quintillion,* $2.5\times10^{18}$).

Si bien sabemos que existe una amplia variedad de tipos de datos a analizar, una buena clasificación nos ayudaría a entender mejor su representación, aunque es muy probable que estas categorías puedan extenderse con el avance tecnológico.

1. *Web and Social Media*: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc, blogs.

2. *Machine-to-Machine (M2M)*: M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3. *Big Transaction Data*: Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4. *Biometrics*: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5. *Human Generated*: Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

En 2001, en un informe de investigación que se fundamentaba en congresos y present aciones relacionadas el analista Doug Laney del META Group (ahora Gartner) definía el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen, la velocidad y la variedad. Gartner continúa usando big data como referencia de este. Además, grandes proveedores del mercado de big data están desarrollando soluciones para atender las demandas más críticas de procesamiento de datos masivos, como MapR, Cyttek Group y Cloudera.

### 1.2.3 La reputación corporativa en los social media

La reputación corporativa en los social media es la percepción hacia una persona o marca en internet, el cual es fabricado por los usuarios cuando conversan y aportan sus opiniones a través de foros, blogs o redes sociales.

Al ser un contenido fácilmente accesible, la reputación es construida desde una multiplicidad de fuentes y ser usada por otros usuarios para realizar fuentes de valor, lo que antes se quedaba en un entorno reducido ahora gracias a intenet la información se distribuye de forma masiva y puede alcanzar grandes cotas mediáticas.

La reputación que vierten los usuarios en la red suele ser muy transcedental para las

ventas de una empresa, ya que estas opiniones se convierten en recomendaciones de personas en los que los usuarios confían al proceder de un consumidor, no como los mensajes publicitarios en masa.

Un aspecto esencial de la gestión de la reputación online es la monitorización en tiempo real de Internet para estar al tanto de ataques potenciales contra la reputación del individuo o de la organización en cuestión. A tal fin las agencias especializadas emplean sofisticadas herramientas tecnológicas que permiten conocer de forma casi instantánea todos los contenidos generados tanto en buscadores como en blogs, redes sociales, etc.

En España, la creciente importancia del control de la reputación online se ha visto reflejada en la creación de departamentos específicos por parte de las agencias más importantes de marketing en buscadores y analítica web para la gestión de la reputación y la monitorización online.


Figura 2. Reputación corporativa

1.2.4 Análisis de sentimiento

El análisis de sentimiento (también conocida como la minería opinión) se refiere al procesamiento del lenguaje natural en textos para identificar y extraer información subjetiva sobre un concepto (empresa, producto, etc).

En términos generales, analizar el sentimiento de un texto tiene como objetivo determinar la actitud de un orador o escritor con respecto a algún tema. El sentimiento de un texto está determinado por varios factores, como el estado emocional del autor o el efecto emocional que el autor desea tener en el lector.

Los primeros trabajos en este campo han sido realizados por Turney y Pang que aplican diferentes métodos para determinar el sentimiento (positivo o negativo) en las opiniones y reseñas sobre películas. También se puede clasificar la polaridad de un documento en una escala de varias vías, es decir, medir la intensidad del sentimiento. A pesar de que en la mayoría de los métodos de clasificación

estadística, la clase neutral se ignora bajo el supuesto de que los textos neutros se encuentran cerca de los límites del clasificador binario. Varios investigadores han demostrado que los clasificadores pueden beneficiarse de la introducción de la clase neutra y mejorar la precisión global de la clasificación.

Otra línea de investigación es la identificación de la subjetividad / objetividad de un texto. Esta tarea es comúnmente define como la clasificación de un texto determinado (por lo general una frase) en una de las dos clases: objetiva o subjetiva. Pang mostró que la eliminación de frases objetivas de un documento antes de clasificar a su polaridad ayudó a mejorar el rendimiento.

La precisión de un sistema de análisis de sentimiento es, en principio, de lo bien que está de acuerdo con los juicios humanos. Según la investigación las personas generalmente están de acuerdo en un 79% de asuntos. Por lo tanto, un clasificador semántico que evalúa correctamente alrededor de un 70% hace tan buen trabajo como un ser humano. Existen medidas más sofisticadas para evaluar un clasificador, por ejemplo si estamos trabajando con escalas. La correlación (distancia entre la clase objetivo y la predicha) podría ser una buena manera de evaluar si nuestro sistema clasifica de una manera correcta.

### 1.2.5 Herramientas para medir reputación corporativa

A continuación se presenta de manera breve las principales herramientas que hay en el mercado para calcular la RC.

#### 1.2.5.1 Reputation Institute: RepTrak®

Reputation Institute es la consultora líder mundial especializada en reputación, fundada en 1997 por el Dr. Charles Fombrun y el Dr. Cees van Riel que colabora con los líderes empresariales para facilitar la tomar decisiones de negocio que construyan y protejan su capital reputacional y conduzcan a la obtención de una ventaja competitiva.

La herramienta creada por el Reputation Institute para medir la RC está basada en el modelo RepTrak®, utilizado para analizar las reputaciones tanto de empresas como de instituciones. Reputation Institute dispone de oficinas y asociados en 30 países del mundo.


Figura 3. Reputation Institute y RepTrak®

RepTrak® evalúa la reputación corporativa de una empresa a través de cuatro atributos (admiración, estima, impresión y confianza) justificadas a través de una serie de dimensiones que define el RC. Estas dimensiones son : (1) Oferta de Productos/ Servicios, (2) Innovación, (3) Entorno de trabajo, (4) Ciudadanía, (5)

Integridad, (6) Liderazgo y (7) Finanzas. Cada dimensión incluye atributos específicos que deben ser hechos a la medida para cada empresa.

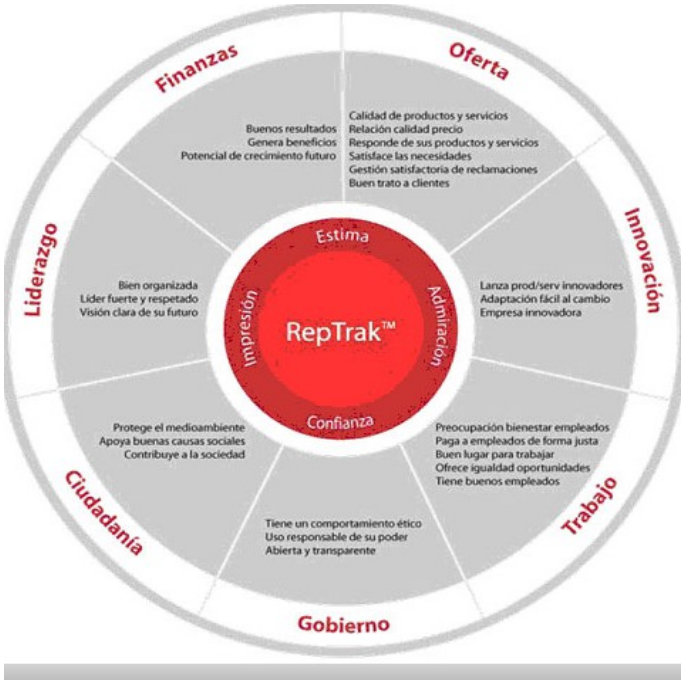
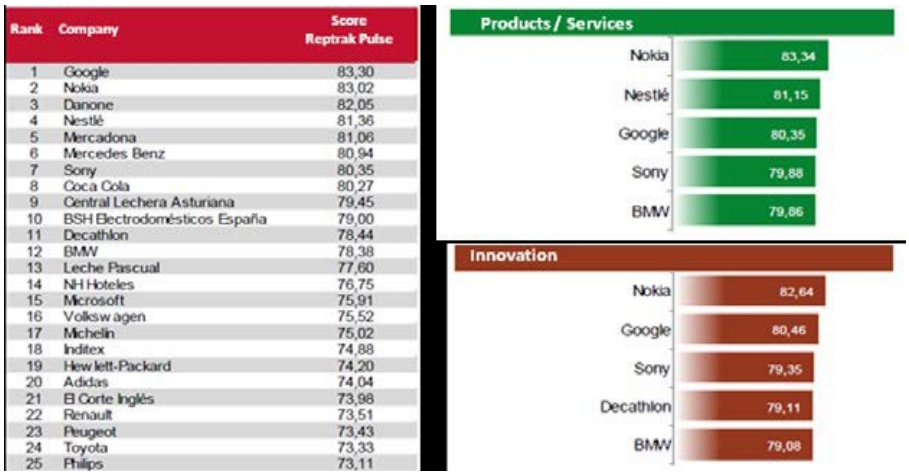



Figura 4. Modelo Reptrak®

Para evaluar cada dimensión (del Reptrak™) se realizan encuestas on line", aunque esto puede cambiar dependediendo de las circunstancias que concurren en cada país o de las peculiaridades del stakeholder que se contempla en el análisis. Por ejemplo, en el caso de Perú, la encuesta es personal y no online, y fue realizado por Inmark Perú. En otros países, como los del G8 (Grupo de los 8 países más industrializados del mundo), donde la utilización de Internet es muy elevada y existen portales dedicados en exclusiva a la realización de estudios de mercado. Hay otros mercados en los que la recogida de datos se lleva a cabo a través de la encuesta telefónica.



| Rank | Company | Score Reptrak Pulse |
|---|---|---|
| 1 | Google | 83,30 |
| 2 | Nokia | 83,02 |
| 3 | Danone | 82,05 |
| 4 | Nestlé | 81,36 |
| 5 | Mercadona | 81,06 |
| 6 | Mercedes Benz | 80,94 |
| 7 | Sony | 80,35 |
| 8 | Coca Cola | 80,27 |
| 9 | Central Lechera Asturiana | 79,45 |
| 10 | BSH Electrodomésticos España | 79,00 |
| 11 | Decathlon | 78,44 |
| 12 | BMW | 78,38 |
| 13 | Leche Pascual | 77,60 |
| 14 | NH Hoteles | 76,75 |
| 15 | Microsoft | 75,91 |
| 16 | Volksw agen | 75,52 |
| 17 | Michelin | 75,02 |
| 18 | Inditex | 74,88 |
| 19 | Hew lett-Packard | 74,20 |
| 20 | Adidas | 74,04 |
| 21 | El Corte Inglés | 73,98 |
| 22 | Renault | 73,51 |
| 23 | Peugeot | 73,43 |
| 24 | Toyota | 73,33 |
| 25 | Philips | 73,11 |

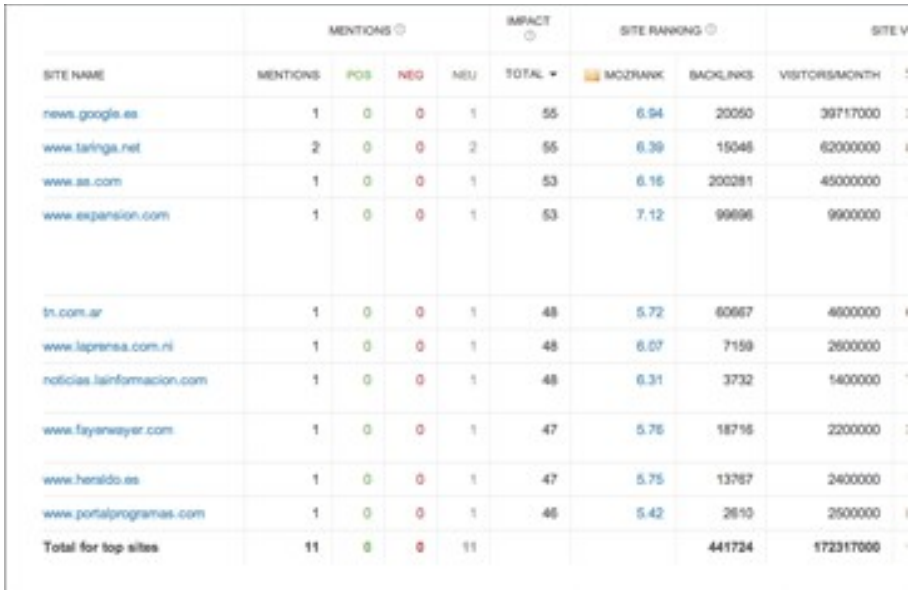


Figura 5. Datos RepTrak

1.2.5.2 BuzzMometer

Es una herramienta que permite monitorizar y analizar la actividad de cualquier marca en la web, desde social media (twitter, facebook) como diarios online, blogs, foros, etc.

Algunas de sus características:
•Rastrea información de más de 70 millones de fuentes
•Permite clasificar la información por fuente, región y género
•Analiza el sentimiento de la información recogida



| | MENTIONS | | | | IMPACT | SITE RANKING | | SITE V |
|---|---|---|---|---|---|---|---|---|
| SITE NAME | MENTIONS | POS | NEG | NEU | TOTAL | MOZRANK | BACKLINKS | VISITORS/MONTH |
| news.google.es | 1 | 0 | 0 | 1 | 55 | 6.94 | 20050 | 39717000 |
| www.taringa.net | 2 | 0 | 0 | 2 | 55 | 6.39 | 15046 | 62000000 |
| www.as.com | 1 | 0 | 0 | 1 | 53 | 6.16 | 200281 | 45000000 |
| www.expansion.com | 1 | 0 | 0 | 1 | 53 | 7.12 | 99696 | 9900000 |
| tn.com.ar | 1 | 0 | 0 | 1 | 48 | 5.72 | 60667 | 4600000 |
| www.laprensa.com.ni | 1 | 0 | 0 | 1 | 48 | 6.07 | 7159 | 2600000 |
| noticias.lainformacion.com | 1 | 0 | 0 | 1 | 48 | 6.31 | 3732 | 1400000 |
| www.fayerwayer.com | 1 | 0 | 0 | 1 | 47 | 5.76 | 18716 | 2200000 |
| www.heraldo.es | 1 | 0 | 0 | 1 | 47 | 5.75 | 13767 | 2400000 |
| www.portalprogramas.com | 1 | 0 | 0 | 1 | 46 | 5.42 | 2610 | 2500000 |
| **Total for top sites** | **11** | **0** | **0** | **11** | | | **441724** | **172317000** |

Figura 6. BuzzMometer. Rastreo de la información

1.2.5.3 Empresas mundiales más admiradas

Desde el año 1987 la consultora Hay Group y la revista Fortune han publicado un ranking de las empresas más admiradas del mundo, que se agrupan en sectores. Esta clasificación se aplica a las empresas con ingresos de un mínimo de 8.000 millones de dólares y la muestra es creado por 10.000 directores y analistas de la compañía. Las dimensiones evaluadas son: la innovación, la calidad de la gestión, el valor de la inversión a largo plazo, la RSE (responsabilidad social empresarial) en la comunidad y en sus alrededores, la capacidad de atraer talento, el servicio y la calidad del producto, la estabilidad financiera, el uso inteligente de los activos y la capacidad de hacer negocios a nivel mundial .

1.2.5.4 Review 200

Este monitor se ocupa de las principales compañías asiáticas y ha sido desarrollado desde 1993 por Far Eastern Economic Review, con la colaboración de DHL Worldwide y AC Nielsen Internacional (Hong Kong). La metodología empleada es una encuesta de opinión / encuesta de los lectores de Far Eastern Economic Review y los editores de las cinco revistas de negocios más importantes de este continente. Los atributos RC medidos son: el servicio al cliente, producto y servicio de calidad, salud financiera, valor de inversión a largo plazo y la innovación.

### 1.2.5.5 Worlds Most Respected Companies

Este ranking fue publicado por primera vez en 1998 y ofrece dos tipos de información: en la primera Detallan las compañías más respetadas y en la segunda se examinan los más respetados jefes de empresa. Los atributos por los que se mide la reputación corporativa son: la creación de valor, la integridad y la responsabilidad social corporativa (RSC).

### 1.2.5.6 Merco empresas

Este monitor, fue desarrollado por la firma de consultoría Villafañe y Asociados, que ofrece información acerca de la reputación de las empresas en España y América del Sur en un nivel global y local. La Reputación Corporativa se mide con base en seis dimensiones: resultados económico-financieros, la calidad del producto-servicio, la cultura corporativa y calidad laboral, ética y responsabilidad social, dimensión global y presencia internacional e innovación. En la siguiente figura se pueden observar las mejores empresas según 7 tipos de público o stakeholders (directores y analistas, sindicatos, organizaciones no gubernamentales, medios de comunicación, líderes de opinión, expertos en materia de RSE y los consumidores).

| Company | 2011 | 2010 | Evolution | Scores |
|---|---|---|---|---|
| Santander | 1º | 2º | ↑ | 10.000 |
| INDITEX | 2º | 3º | ↑ | 9.405 |
| Telefónica | 3º | 1º | ↓ | 9.255 |
| REPSOL | 4º | 6º | ↑ | 8.942 |
| "la Caixa" | 5º | 4º | ↓ | 8.885 |
| MERCADONA | 6º | 9º | ↑ | 8.878 |
| IBERDROLA | 7º | 5º | ↓ | 8.626 |
| BBVA | 8º | 7º | ↓ | 8.064 |
| El Corte Inglés | 9º | 8º | ↓ | 7.554 |
| MAPFRE | 10º | 10º | = | 7.269 |



Figura 7. Merco Empresas

### 1.2.6 Herramientas para gestionar la reputación corporativa

Las empresas, además de alimentarse de la información obtenida anualmente de los monitores antes mencionados, requieren herramientas que faciliten la toma de decisiones al establecer políticas y estrategias de RC centradas en cada uno de sus públicos estratégicos. Este párrafo se refiere a las principales herramientas utilizadas para la gestión de la reputación corporativa. Ellos son: RepTrak, TRI * M y Alva. Todos ellos han sido desarrollados por empresas de consultoría con una amplia experiencia en el estudio de la reputación corporativa Además, un estudio comparativo se llevó a cabo para comprender cuáles son sus fortalezas y debilidades en la actualidad.

### 1.2.6.1 TRI * M

Desarrollado por la consultora TNS, TRI * M (O'Gorman y Pirner, 2006) es una herramienta que se utiliza para apoyar la reputación en términos de medición, gestión y supervisión o el control de las relaciones con los públicos estratégicos de la compañía. La evaluación de las tendencias a largo plazo de las diferentes

unidades de negocio y la identificación de los puntos fuertes en las relaciones con cada uno de sus públicos o stakeholders.

TRI * M tiene a su disposición una base de datos de más de 1.000 empresas en todo el mundo, que les permite hacer un análisis comparativo e interpreta los resultados de cada estudio. Con más de 6.000 estudios y más de 10 millones de entrevistas que permite la identificación del público al que la empresa debe dirigir sus acciones para mejorar su RC.

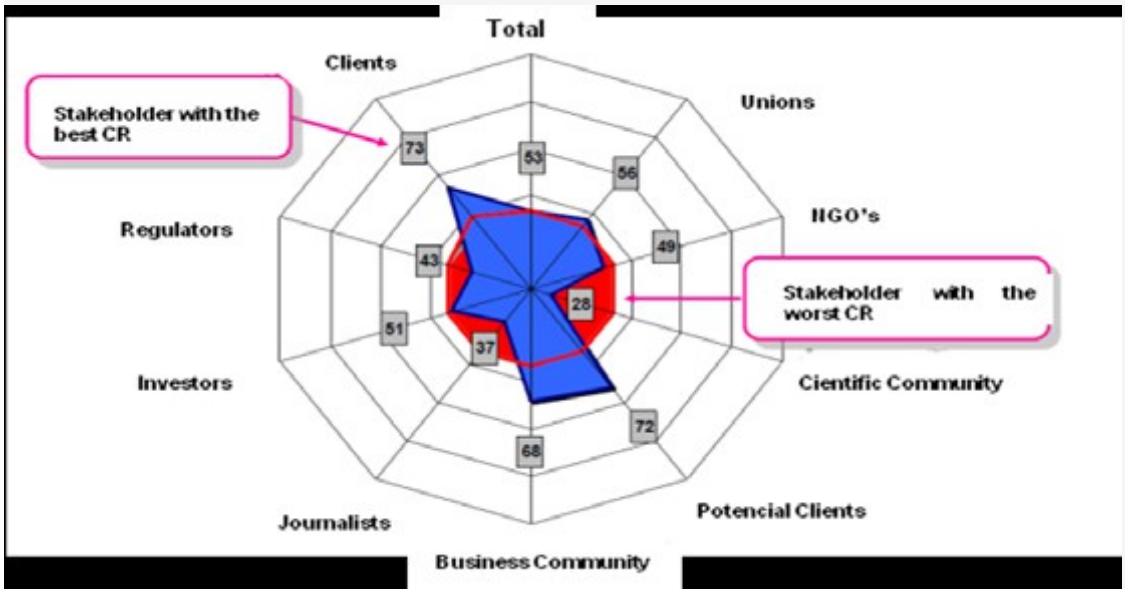



Figura 8. TRI * M Stakeholders

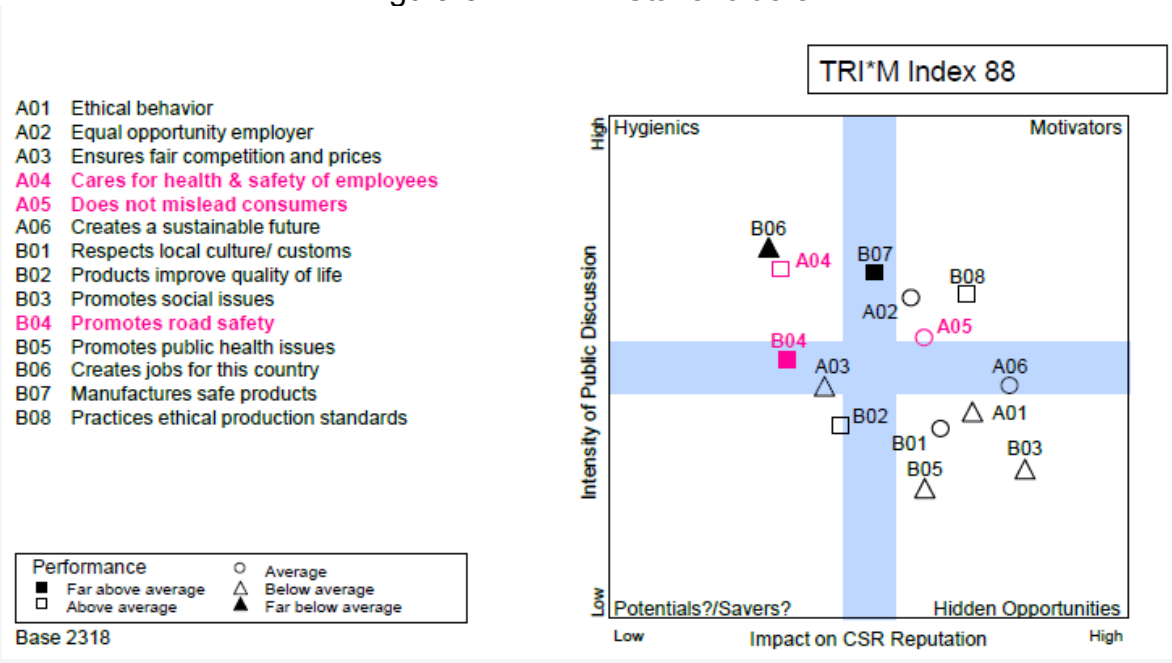



Figura 9. TRI * M Index 88

### 1.2.6.2 Alva in Depth

Desarrollado por la empresa de consultoría Alva (Alva, 2011), es una herramienta para gestionar la reputación de empresa; que permite la medición de la RC, ver que asuntos afectan actualmente a la empresa, la comparación de la RC de la empresa con su sector ya sea a nivel mundial, nacional o regional.


Figura 10. Alva in Depth

Esta herramienta establece reputación de la empresa mediante el análisis de la misma, las fuentes y los sectores de la reputación.

El análisis se lleva a cabo sobre una base global utilizando fuentes más allá de los medios de comunicación, dando una amplia comprensión de las percepciones que tienen los stakeholders. Cerca de un millón de piezas de contenido se procesan diariamente y se actualiza cada 60 segundos. Cuenta con un gran número de fuentes a su disposición; social media, los medios de comunicación tradicionales, publicidad, análisis de expertos, informes de agentes, analistas financieros, encuestas de opinión pública, estudios, datos.

También permite analizar la reputación corporativa de un sector, por lo que permite a las organizaciones comparar su reputación con la de sus competidores, contextualizando así su comportamiento y resultados.

Por otra parte, esta herramienta proporciona a la compañía una comprensión más profunda de las siguientes áreas con el fin de gestionar su reputación:

1. Mediante el análisis de una serie de atributos, los aspectos del desempeño de la empresa que afectan a su reputación son identificados, así como las zonas de las que proceden y de las medidas necesarias que deben adoptarse con el fin de mejorar el rendimiento.

2. La herramienta permite la gestión de los temas de riesgo ante una crisis corporativa.

3. El sistema muestra los puntos de reputación individual de cada empresa y por sectores. En la figura se pueden observar las tendencias y los movimientos del sector y de la empresa de una manera más estructurada mediante el seguimiento de fuentes distintas.


Figura 11.Analisis reputacional de una compañía. Fuente: Alva

## 1.3. Estructura de la memoria

Dividimos la memoria en varios apartados:

- Un primer apartado donde se habla nuestra propuesta y lo que vamos a desarrollar, explicando cada uno de los distintos apartados que compone nuestro proyecto
- Luego en el apartado de implementación se explica las herramientas de desarrollo que se han utilizado, librerías y algunos detalles más. También se explicará la base de datos así como el modelo entidad-relación usado como base de la aplicación.
- Conclusiones donde se destaca todo lo que hemos aprendido en el desarrollo del TFG, así como algunas pautas de por donde debería seguir en un futuro.

## 2. Especificación y requisitos

### 2.1 Nuestra propuesta

Nuestra propuesta es desarrollar un sistema que se encargue de calcular la reputación corporativa de una empresa en internet. Para calcular la RC necesitamos tener a disposición un conjunto de datos, los cuales el sistema también se encargará de obtenerlo.

Para facilitar la tarea hemos dividido nuestro sistema en tres módulos/niveles que explicaremos a continuación:

- Captura de información: Se encarga de capturar información de las empresas a través de la web, para luego ser guardada en una base de datos.

- Tratamiento de la información: Haciendo uso de la información guardada en la base de datos, aplicamos un clasificador semántico que se va a encargar de evaluar el sentimiento de cada información para luego poder calcular la reputación de cada empresa con la información analizada.

- Visualización de la información: Crear una interfaz gráfica para visualizar el resultado obtenido.

Estos módulos son independientes entre sí, cuyo único punto en común es la conexión a la misma base de datos, por lo que en la memoria dedicaremos un capítulo exclusivo explicando la implementación para cada uno de estos módulos.

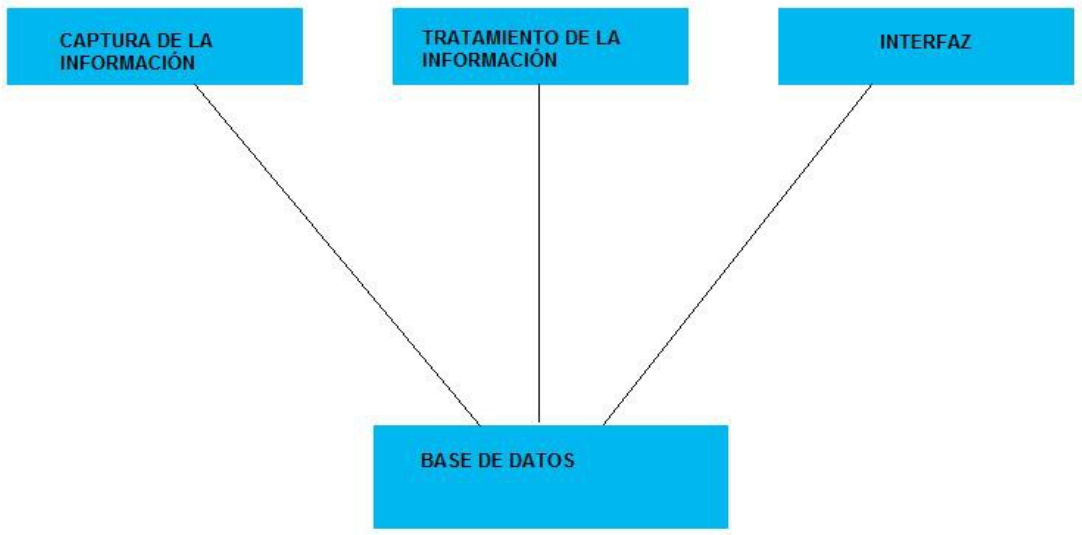



Figura 12. Esquema sistema

#### 2.1.1 Captura de información

Este módulo es el que se encargará nutrir nuestro sistema de la información necesaria su funcionamiento.

Hemos estado barajando distintos social media desde donde obtener la información que nutrirá a nuestro sistema y nos hemos quedado con la prensa digital por las

siguientes razones:

- Facilidad para recoger datos: la mayoría de los diarios tienen implementado en sus webs RSS para compartir el contenido web, por lo que facilita la recolección automática de la información sin tener el cuenta el diseño de la página web.

- Analisis semántico: La forma de escribir en estos medios y en la redes sociales son distintas por el tipo de usuario que escribe el mensaje. Mientras que en redes sociales se suele limitar mucho el lenguaje (faltas ortográficas, limitación en el número de palabras, etc) en diarios y blogs se suele escribir de una manera correcta, que facilita luego el análisis del texto.

Por esto es que nos centraremos en los diarios digitales que ofrezcan el formato RSS para difundir la información. El RSS da toda la información de cada uno de las noticias que son publicadas en su web, así como de su título, cabecera, fecha de publicación, autor, etc…

Nuestro sistema se encargará de filtrar los datos obtenidos de la lectura de RSS para obtener sólo las noticias que contengan en su título el nombre de alguna de las empresas guardadas en la base de datos y de la que queremos obtener su RC. Por ejemplo, si tenemos guardada en la base de datos la empresa Telefónica, una posible noticia podría ser: “La Fundación Santa María la Real y Telefónica unen fuerzas en la conservación del patrimonio”.

#### 2.1.1.2 RSS

RSS son las siglas de Really Simple Syndication, un formato XML para sindicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos. El formato permite distribuir contenidos sin necesidad de un navegador, utilizando un software diseñado para leer estos contenidos RSS tales como Internet Explorer, entre otros (agregador). A pesar de eso, es posible utilizar el mismo navegador para ver los contenidos RSS. Las últimas versiones de los principales navegadores permiten leer los RSS sin necesidad de software adicional. RSS es parte de la familia de los formatos XML, desarrollado específicamente para todo tipo de sitios que se actualicen con frecuencia y por medio del cual se puede compartir la información y usarla en otros sitios web o programas. A esto se le conoce como redifusión web o sindicación web (una traducción incorrecta, pero de uso muy común).


Figura 13. Ejemplo RSS

### 2.1.2 Tratamiento de la información (análisis semántico basado en expresiones)

Nuestro clasificador está basado en analizar en sentimiento del texto hacia un concepto, en nuestro caso ese concepto sería la que queremos valorar. Nuestro objetivo es valorar el sentimiento de las palabras que aparecen en el texto y que están relacionadas con la empresa.

El primer paso es encontrar el concepto que queremos analizar. El primer módulo del sistema consistió en la búsqueda y captura de información en el que apareciera alguna de las empresas. En la base de datos tenemos asociado a que empresas está asociada cada información, por lo que el concepto al cual queremos analizar ya lo tenemos.

El siguiente paso consiste en valorar las palabras que están relacionadas con la empresa, a este concepto le hemos dado el término de expresiones.

El concepto de expresión es el mismo que el de expresión regular, es decir una secuencia de caracteres que forma un patrón de busqueda que se buscan dentro de un texto para analizar su sentimiento. Ejemplo de una expresión <ExprCOMPRAR>.*<AdjetivosBUENOS> le hemos dado tono positivo (+1). <ExprCOMPRAR> y <AdjetivosBUENOS> son etiquetas las cuales explicaremos a continuación.

Una etiqueta está formada por una serie de palabras asociados a una clase, como puede ser verbos positivos, adjetivos negativos, etc que ayudan a la construcción de las expresiones y se usan para facilitar el entendimiento de las expresiones. Por ejemplo para <AdjetivosBUENOS> la expresión asociada es: (renueva|renova[rc]| innova|mejora|moderniza|actualiza|descubr[ei]|inventa|perfecciona|reforma|progresa| inven[tc]|optimiza), es decir, un conjunto de adjetivos positivos.

Ahora vamos a explicar el algoritmo para analizar un texto:

- Realizamos un preprocesado del texto (quitar links, letras repetidas, tildes)
- Luego se procede a contar el nº de empresas, expresiones encontradas y negaciones.
- Si es un caso válido:
  - Ver la figura 14.
- En otro caso
  - Dividir el texto en frases (usando los signos de puntuación)
  - Repetir desde el punto 2.
  - La valoración del texto es la suma de las valoraciones de todas las frases.

| Negaciones | Empresas | Expresiones | Valoración |
|---|---|---|---|
| 0 | 1 | 1 | Tono de la expresión |
| 0 | >1 | 1 | A todas la empresas se le asigna la empresa del texto |
| 0 | 1 | >1 | A la empresa se le asigna la suma del tono de todas las expresiones |
| Otro caso | | | Implementación en un futuro |

Figura 14. Tabla valoraciones

Si la valoración de un texto es mayor que 2, se deja en 2, para así evitar que el peso de un texto sea demasiado grande a la hora de calcular la reputación de una empresa, lo mismo para una valoración negativa (si es menor de -2 lo dejamos en -2).

Para Calcular el RC de una empresa simplemente calculamos la media aritmética de todas las expresiones donde aparece la empresa o uno de sus sinónimos.

En la siguiente tabla damos el significado del RC según su puntuación:

| Intervalo | Significado |
|---|---|
| [-2,-1] | Muy negativo |
| [-1,0] | Negativo |
| [0] | Neutro |
| [0,1] | Positivo |
| [1,2] | Muy positivo |

Figura 15. Significado RC

2.1.3 Visualización de la información
Este módulo está formado por una pequeña interfaz que muestra los resultados obtenidos por los dos módulos anteriores.

La interfaz estará compuesto por varias pestañas:

Una primera pantalla con un menú desplegable con distintas empresas que al seleccionar cualquiera de ellas podemos ver su reputación corporativa.

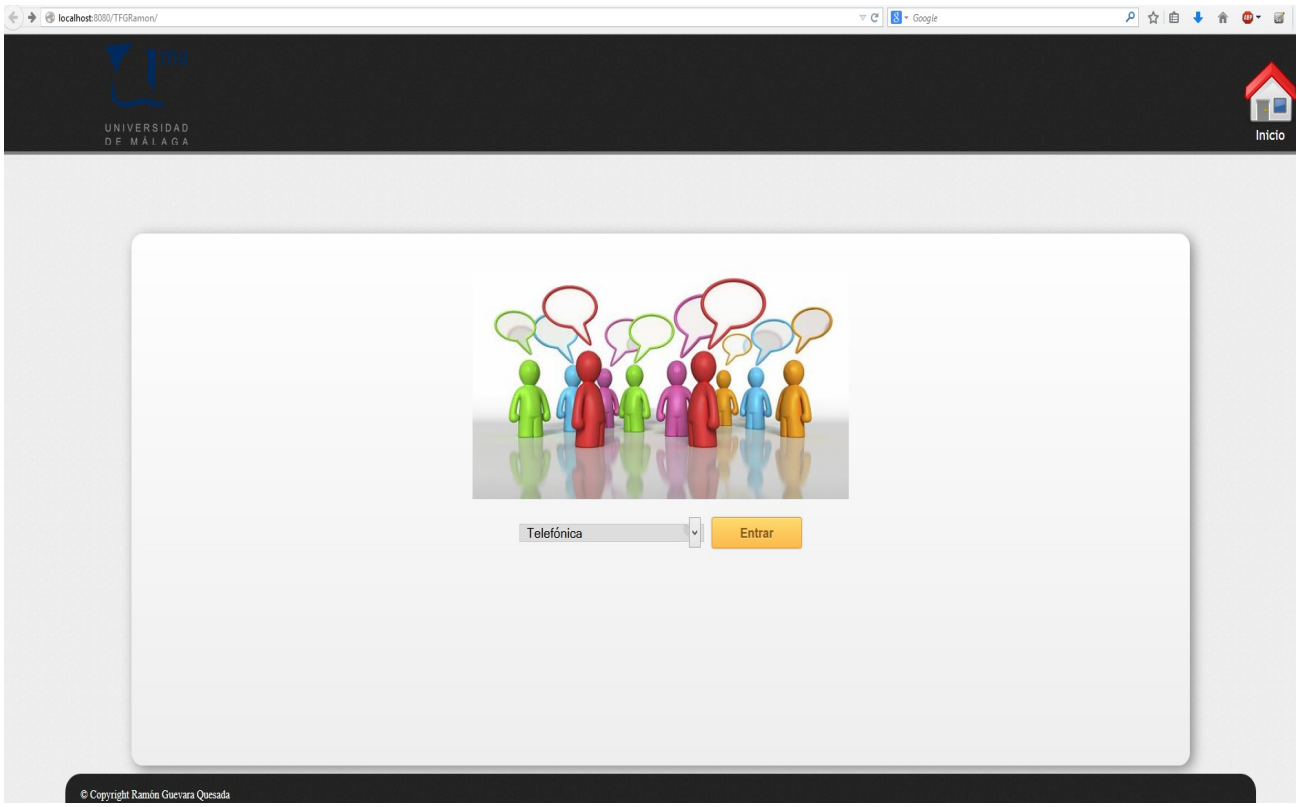



Figura 16. Pantalla principal interfaz

En la segunda pantalla se muestra dos gráficos. El primero muestra el número de noticias son positivas, es decir, que hablan bien de la empresa seleccionada y el número de noticias negativas. El segundo gráfica muestra la RC de la empresa.

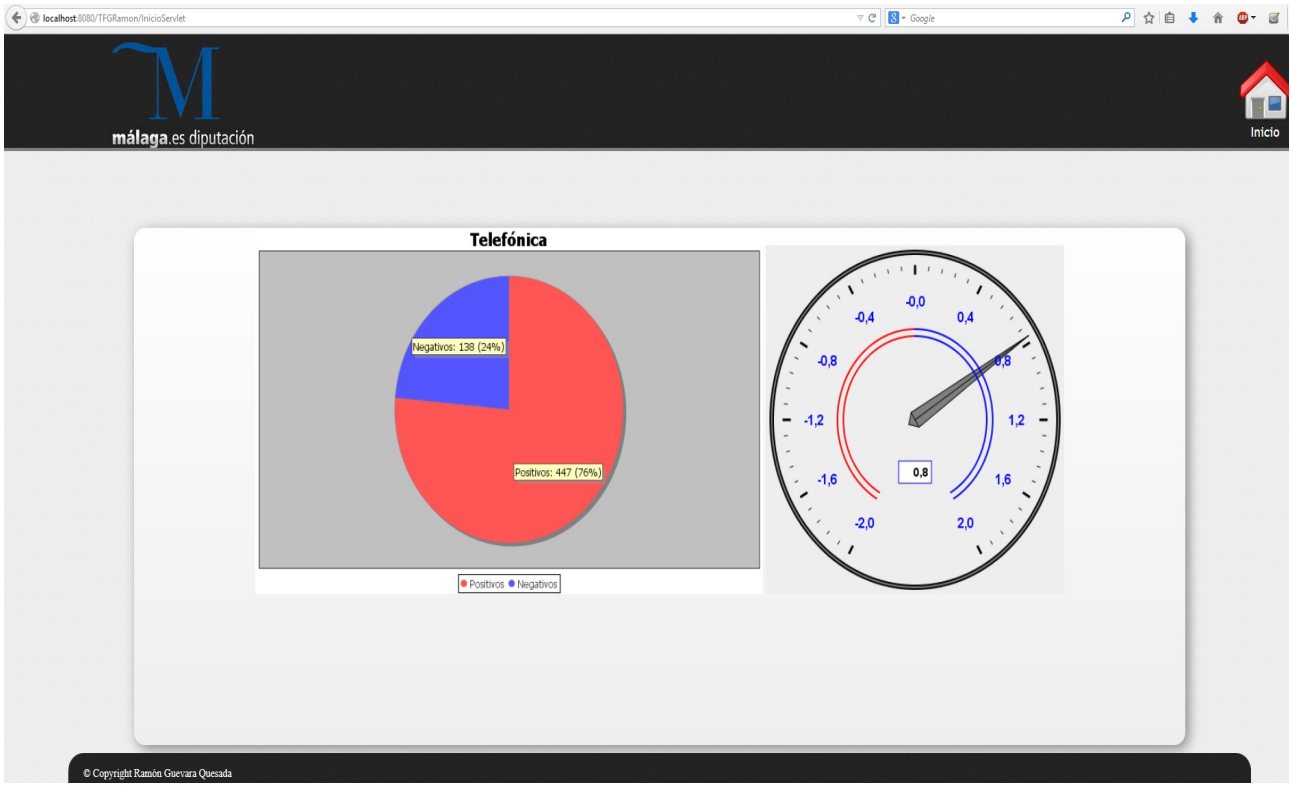



Figura 17. Pantalla reputación interfaz

2.2 Requisitos funcionales:

◦Lectura de medios: Una vez que se seleccionen los medios que se van a leer, el sistema debe ser capaz de capaz de realizar una lectura cada cierto tiempo para obtener toda la información que haga referencia a las empresas seleccionadas.

◦Calculo de RC: Para ello tenemos que evaluar toda la información recogida, es decir aplicar el clasificador semántico que evalúa el tono del texto y luego aplicar otro algoritmo que dada las evaluaciones que hay de una empresa, calcule su reputación corporativa.

◦Selección de empresa: Elegir la empresa de la cual queremos saber su RC.

◦Visualizar en pantalla la RC de la empresa seleccionada.

◦Aparte de visualizar el RC de la empresa, tener acceso a la información capturada, pudiendo la información obtenida dentro de un rango obtenido o si solo se quiere mostrar la información de un medio en particular.

2.3 Requisitos no funcionales

◦Intuitiva: Queremos que nuestra interfaz sea fácil de utilizar para cualquier tipo de usuario, aunque no tenga muchos conocimientos de informática.

◦Consistencias y estándares: gran distinción entre todas las acciones que se puedan realizar y sus nombres para evitar confusiones.

◦Prevención de errores: Garantizar que la aplicación y los datos de entrada sean robustos y consistentes.

◦Diseño minimalista : Que la interfaz sólo tenga los datos necesarios para su uso y nada más.

## 3. Implementación

Ahora explicaremos como implementar cada uno de los tres módulos que componen nuestro sistema. Empezaremos explicando primero el modelo de base de datos que hemos construido, ya que es común para toda la aplicación y luego cada uno de los módulos inpependientemente.

### 3.1 Base de Datos

Microsoft SQL Server 2012 es un sistema para la gestión de bases de datos producido por Microsoft basado en el modelo relacional. Sus lenguajes para consultas son T-SQL y ANSI SQL. Microsoft SQL Server constituye la alternativa de Microsoft a otros potentes sistemas gestores de bases de datos como son Oracle, PostgreSQL o MySQL.

#### 3.1.1 Modelo Entidad-Relación

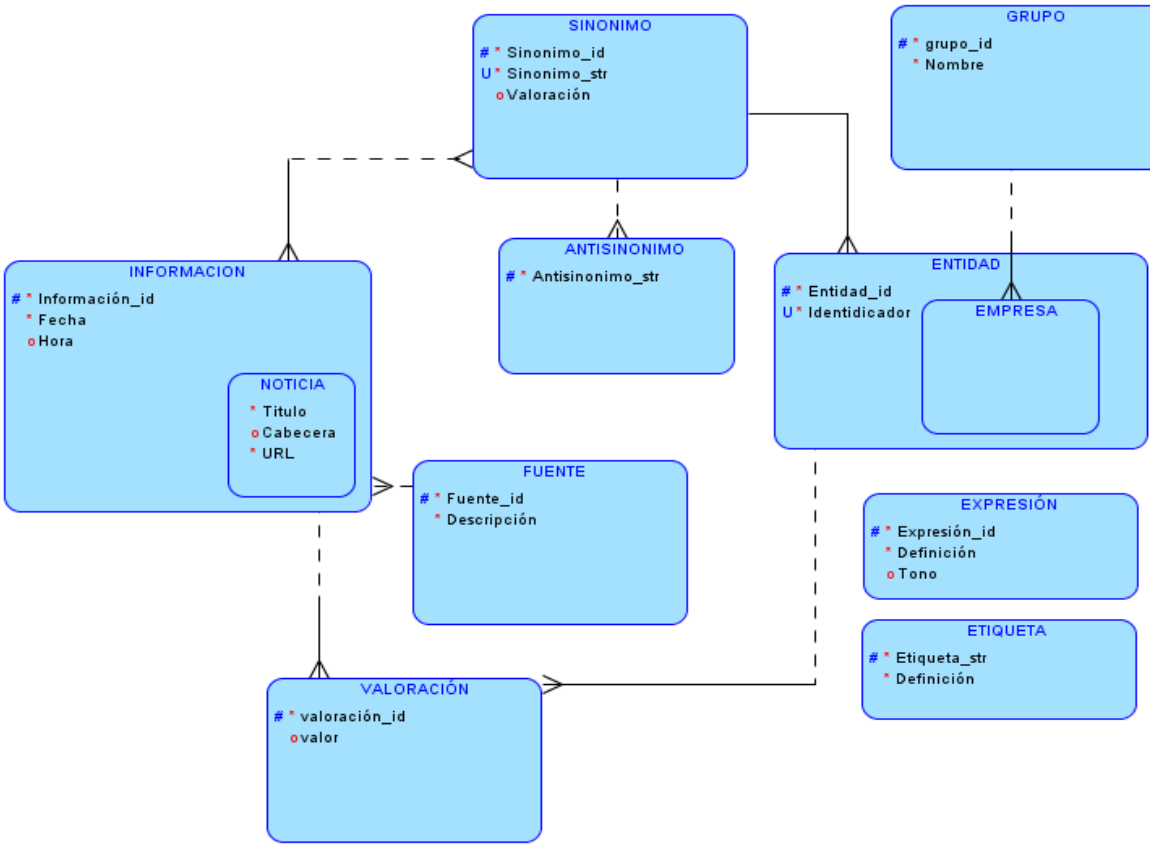



Figura 18. Modelo Entidad-Relación

Ahora pasemos a explicar cada una de las tablas:

| **ENT – 01** | **INFORMACIÓN** | | |
|---|---|---|---|
| **Descripción** | Llamaremos INFORMACIÓN a toda porción de texto sin procesar por el sistema que se ha leído como una unidad de alguna de las fuentes consideradas. Puede ser de varios tipos en función de la fuente de donde se ha extraído. | | |
| **Supertipos** | | | |
| **Subtipos** | NOTICIA (ENT-02) | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |
| **Atributos** | ATR–01: Información_id<br>ATR–02: Fecha<br>ATR–03: Hora | | |
| **Comentarios** | Ninguno | | |

| **ATR – 01** | **INFORMACIÓN:: Información_id** |
|---|---|
| **Descripción** | Identificador único para cada porción de información |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria, auto-incremental |

| **ATR – 02** | **INFORMACIÓN:: Fecha** |
|---|---|
| **Descripción** | Indica la fecha de publicación de la información según la fuente de donde se ha leído |
| **Tipo** | Fecha |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| **ATR – 03** | **INFORMACIÓN:: Hora** |
|---|---|
| **Descripción** | Indica la hora de publicación de la información según el medio de donde se ha leído |
| **Tipo** | Hora |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Puede ser nulo |

| | |
|---|---|
| | En caso de que el medio no ofrezca la hora de publicación podríamos usar la hora en la que el sistema ha leído la información (siempre que la fecha de publicación sea la misma que la fecha de lectura) |

| **ENT – 02** | **NOTICIA** | | |
|---|---|---|---|
| **Descripción** | Una NOTCIA es una texto con información leída de la página web de un medio de comunicación. Accedemos a las noticias a través de la lista de noticias publicadas a través de la RSS de cada periódico. | | |
| **Supertipos** | INFORMACIÓN (ENT-01) | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |
| **Atributos** | ATR–04: Título<br>ATR–05: Cabecera<br>ATR–06: URL | | |
| **Comentarios** | Hereda la clave primaria de INFORMACIÓN | | |

| **ATR – 04** | **NOTICIA:: Título** |
|---|---|
| **Descripción** | Almacena el TÍTULO de la NOTICIA |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| **ATR – 05** | **NOTICIA:: Cabecera** |
|---|---|
| **Descripción** | Almacena la CABECERA de la NOTICIA |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Puede ser nulo<br>Es un texto que acompaña a veces en cuerpo de la NOTICIA a modo de resumen |

| **ATR – 06** | **NOTICIA:: URL** |
|---|---|
| **Descripción** | Es la dirección web exacta de dónde se ha leído la NOTICIA |
| **Tipo** | Cadena de caracteres |

| | |
|---|---|
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| **ENT – 03** | **FUENTE** | | |
|---|---|---|---|
| **Descripción** | Son las distintas fuentes desde donde se extrae la información, en nuestro caso los distintos periódicos y blogs (ABC, El País, Menéame, etc) | | |
| **Supertipos** | | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |
| **Atributos** | ATR–07: Fuente_id<br>ATR–08: Descripción | | |
| **Comentarios** | Ninguno | | |

| **ATR – 07** | **FUENTE::Fuente_id** |
|---|---|
| **Descripción** | Identificador único para cada fuente |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria, auto-incremental |

| **ATR – 08** | **FUENTE:: Descripción** |
|---|---|
| **Descripción** | Texto que describe la fuente: en nuestro caso al periódico |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| **ENT – 04** | **ENTIDAD** |
|---|---|
| **Descripción** | Bajo el concepto de ENTIDAD almacenaremos cualquier empresa, persona u organismo, del que bien nos interesa almacenar información, o bien publica información de una entidad que sí nos interesa. Dependiendo de su naturaleza puede ser de distintos tipos |
| **Supertipos** | |
| **Subtipos** | EMPRESA (ENT-05) |

| Componentes | Nombre | Tipo | Mult. |
|---|---|---|---|
| | | | |
| **Atributos** | ATR–09: Entidad_id<br>ATR–10: Descripción | | |
| **Comentarios** | Ninguno | | |

| ATR – 09 | ENTIDAD:: Entidad_id |
|---|---|
| **Descripción** | Identificador único para cada entidad |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria, auto-incremental |

| ATR – 10 | ENTIDAD:: Descripción |
|---|---|
| **Descripción** | Nombre o razón social de la empresa o entidad |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| ENT – 05 | EMPRESA | | |
|---|---|---|---|
| **Descripción** | Una EMPRESA es el tipo de ENTIDAD más importante en nuestro sistema. | | |
| **Supertipos** | ENTIDAD (ENT-04) | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |
| **Atributos** | | | |
| **Comentarios** | Hereda la clave primaria de ENTIDAD | | |

| ENT – 06 | SINÓNIMO | | |
|---|---|---|---|
| **Descripción** | Son cada una de las expresiones alternativas/sinónimos que podemos encontrar en las distintas fuentes de información para referirnos a una misma ENTIDAD. | | |
| **Supertipos** | | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |

| | | | |
|---|---|---|---|
| | | | |
| **Atributos** | ATR–11: Sinónimo_id (clave primaria)<br>ATR–12: Sinónimo<br>ATR–13: Valoración | | |
| **Comentarios** | Es una entidad débil de ENTIDAD (ENT – 04) | | |

| **ATR – 11** | **SINÓNIMO:: Sinonimo_id** |
|---|---|
| **Descripción** | Identificador único para cada sinónimo |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria, auto-incremental |

| **ATR – 12** | **SINÓNIMO:: Sinonimo_str** |
|---|---|
| **Descripción** | Nombre del sinónimo |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| **ATR – 13** | **SINÓNIMO:: Valoración** |
|---|---|
| **Descripción** | Sentimiento del sinónimo. Ejemplo “Vomiestar” es negativo |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | |

| **ENT – 07** | **ANTISINONIMO** | | |
|---|---|---|---|
| **Descripción** | Son expresiones que al incluir el nombre de una entidad, parece que hace referencia a una entidad cuando en realidad no lo está haciendo. Por Ejemplo 'Cabina Telefónica' no está haciendo referencia a la empresa telefónica | | |
| **Supertipos** | | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |

| | |
|---|---|
| **Atributos** | ATR–14: Antisinonimo_str |
| **Comentarios** | Tiene un atributo para hacer referencia a la entidad a la que pertenece el antisinonimo (Entidad_id) |

| **ATR – 14** | **ANTISINONIMO:: Antisinonimo_str** |
|---|---|
| **Descripción** | Nombre del antisinónimo |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria |

| **ENT – 08** | **GRUPO** | | |
|---|---|---|---|
| **Descripción** | Los distintos sectores a los que puede pertenecer una empresa. Por ejemplo telecomunicaciones, construcción, etc. | | |
| **Supertipos** | | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |
| **Atributos** | ATR–15: Grupo_id<br>ATR–16: Nombre | | |
| **Comentarios** | Ninguno | | |

| **ATR – 15** | **GRUPO:: Grupo_id** |
|---|---|
| **Descripción** | Identificador único de cada grupo |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria, autoincremental |

| **ATR – 16** | **GRUPO:: Nombre** |
|---|---|
| **Descripción** | Nombre del sector al que puede pertenecer una entidad |
| **Tipo** | Cadena de caracteres |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| ENT – 09 | VALORACIÓN | | |
|---|---|---|---|
| **Descripción** | Contiene el sentimiento de la INFORMACIÓN una vez procesada | | |
| **Supertipos** | | | |
| **Subtipos** | | | |
| **Componentes** | **Nombre** | **Tipo** | **Mult.** |
| | | | |
| **Atributos** | ATR–17: Valoracion_id<br>ATR–18: Valor<br>ATR-24: Modo | | |
| **Comentarios** | Una valoración hace referencia a una información | | |

| ATR – 17 | VALORACIÓN:: Valoracion_id |
|---|---|
| **Descripción** | Identificador único de cada valoración |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | Clave primaria, autoincremental |

| ATR – 18 | VALORACIÓN:: Valor |
|---|---|
| **Descripción** | Indica de 1 a 10 el sentimiento de la información valorada |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |
| **Comentarios** | No nulo |

| ATR – 24 | VALORACIÓN:: Modo |
|---|---|
| **Descripción** | Indica si es una valoración manual o se ha usdo el método de expresiones. Valores:<br>1: Manual<br>2: Expresiones |
| **Tipo** | Entero |
| **Valor inicial** | |
| **Expresión** | |

| ENT – 10 | EXPRESIÓN | | |
|---|---|---|---|
| Descripción | Contiene expresiones regulares que se buscan dentro de un texto para analizar su sentimiento, el sentimiento de un texto es la suma de todas las expresiones regulares encontradas en un texto. Ejemplo de una expresión <ExprCOMPRAR>.*<AdjetivosBUENOS> con un tono positivo (+1). El significado de <ExprCOMPRAR> y <AdjetivosBUENOS> lo veremos ahora en la tabla etiqueta, es decir, son etiquetas que a su vez son también expresiones regulares. | | |
| Supertipos | | | |
| Subtipos | | | |
| Componentes | Nombre | Tipo | Mult. |
| | | | |
| Atributos | ATR–19: Expresión_id<br>ATR–20: Definición<br>ATR–21: Tono | | |
| Comentarios | Ninguno | | |

| ATR – 19 | EXPRESIÓN:: Expresión_id |
|---|---|
| Descripción | Identificador único de cada expresión |
| Tipo | Entero |
| Valor inicial | |
| Expresión | |
| Comentarios | Clave primaria, autoincremental |

| ATR – 20 | EXPRESIÓN:: Definición |
|---|---|
| Descripción | Definición de la expresión |
| Tipo | Cadena de caracteres |
| Valor inicial | |
| Expresión | |
| Comentarios | No nulo |

| ATR – 21 | EXPRESIÓN:: Tono |
|---|---|
| Descripción | Sentimiento de la expresión |
| Tipo | Entero |
| Valor inicial | |
| Expresión | |
| Comentarios | No nulo |

| ENT – 01 | ETIQUETA | | |
|---|---|---|---|
| Descripción | Son una serie de palabras asociados a una clase, como puede ser verbos positivos, adjetivos negativos, etc que ayudan a la construcción de las expresiones. | | |
| Supertipos | | | |
| Subtipos | | | |
| Componentes | Nombre | Tipo | Mult. |
| | | | |
| Atributos | ATR–22: Etiqueta_str<br>ATR–23: Definicion | | |
| Comentarios | Ninguno | | |

| ATR – 22 | ETIQUETA:: Etiqueta_str |
|---|---|
| Descripción | Representa la clase de la etiqueta (verbos, adjetivos buenos, etc.) |
| Tipo | Cadena de caracteres |
| Valor inicial | |
| Expresión | |
| Comentarios | Clave primaria |

| ATR – 23 | ETIQUETA:: Definicion |
|---|---|
| Descripción | Expresión regular asociada a la etiqueta. Por ejemplo para <AdjetivosBUENOS> la expresión asociada es: (renueva\|renova[rc]\|innova\|mejora\|moderniza\|actualiza\| descubr[ei]\|inventa\|perfecciona\|reforma\|progresa\| inven[tc]\|optimiza), es decir, un conjunto de adjetivos positivos. |
| Tipo | Cadena de caracteres |
| Valor inicial | |
| Expresión | |
| Comentarios | No nulo |

## 3.2 Captura de la información

Comenzaremos explicando cómo hemos construido el módulo de nuestro sistema que se encarga de la recogida de información por las distintas webs. Para eso comencemos explicando las herramientas que hemos usado

### 3.2.1 Herramientas de desarrollo

3.2.1.1 Eclipse

Es el entorno de desarrollo que hemos elegido para desarrollar el código de programación de este módulo, el cual te permite programar usando distintos lenguajes de programación como C, C++, Java, HTML. En nuestro caso hemos usado JAVA, por su gran uso y al ser el lenguaje de programación que más hemos usado a lo largo de la carrera, también nos resulta muy familiar.

Eclipse es un programa informático compuesto por un conjunto de herramientas de programación de código abierto que puede ser descargado desde su página web https://www.eclipse.org/downloads/

De todas las versiones hemos elegido la versión standard la cual tiene todos los paquetes de Java necesarios para realizar el módulo más algunas librerías que le añadiremos y explicaremos a continuación (usaremos la versión 4.2 a diferencia de la que aparece en la imagen que es la 4.4, ya que es la versión que instalamos en su día en nuestro ordenador).

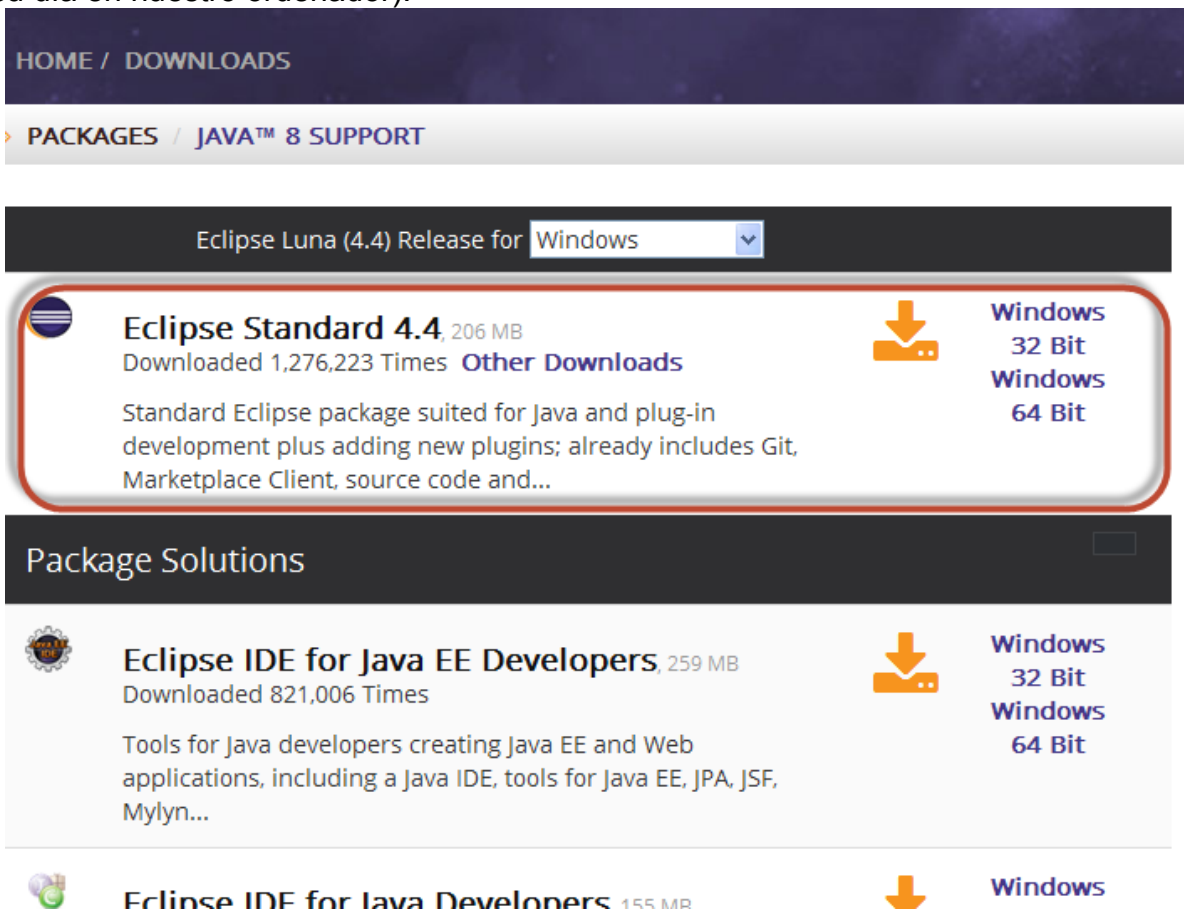



Figura 19. Descarga de eclipse

3.2.1.4 JBoss Tools(Hibernate Tools)

JBoss es un servidor de aplicaciones Java EE de código abierto implementado en Java puro. Al estar basado en Java, JBoss puede ser utilizado en cualquier sistema operativo para el que esté disponible la máquina virtual de Java. JBoss Inc., empresa fundada por Marc Fleury y que desarrolló inicialmente JBoss, fue adquirida por Red Hat en abril del 2006.

Las características destacadas de JBoss incluyen :

- Producto de licencia de código abierto sin coste adicional.
- Cumple los estándares.
- Confiable a nivel de empresa
- Incrustable, orientado a arquitectura de servicios.
- Flexibilidad consistente
- Servicios del middleware para cualquier objeto de Java.
- Soporte completo para JMX.

Para el proyecto se ha usado Jboss Tools, el cual es es un conjunto de plugins para Eclipse que complementa, mejora y va más allá del apoyo que existe para JBoss y las tecnologías relacionadas en la distribución por defecto Eclipse. Uno de los módulos que componen Jboss Tools es Hibernate Tools,

### 3.2.1.3 Hibernate

Hibernate es la librería que hemos elegido para poder operar con la base de datos usando código JAVA.

Hibernate es una herramienta de Mapeo objeto-relacional (ORM) para la plataforma Java (y disponible también para .Net con el nombre de NHibernate) que facilita el mapeo de atributos entre una base de datos relacional tradicional y el modelo de objetos de una aplicación, mediante archivos declarativos (XML) o anotaciones en los beans de las entidades que permiten establecer estas relaciones.

Hibernate es software libre, distribuido bajo los términos de la licencia GNU LGPL, el cual puede ser descargado desde el siguiente enlace [http://hibernate.org/orm/downloads/](http://hibernate.org/orm/downloads/).

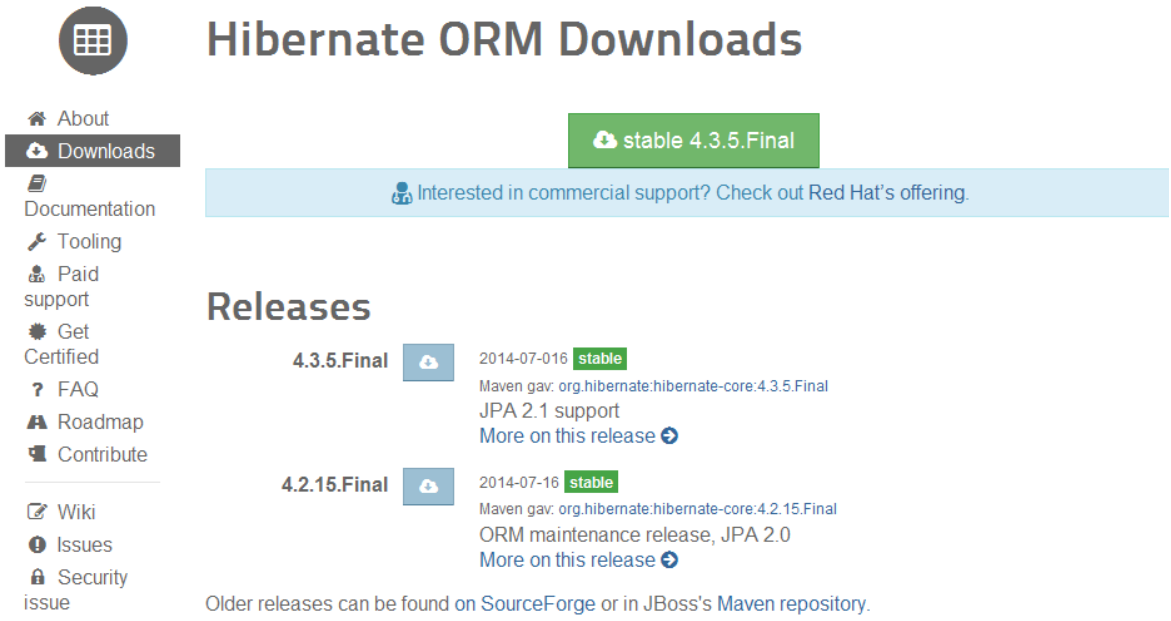



Figura 20. Descarga de hibernate

### 3.2.1.4 ROME

ROME es una librería para Java usada para leer y generar contenido en formato RSS, es un software de código abierto bajo la licencia Apache 2.0. . ROME puede ser descargado desde el siguiente enlace http://rometools.github.io/rome/ROMEReleases/ROME1.0Release.html.

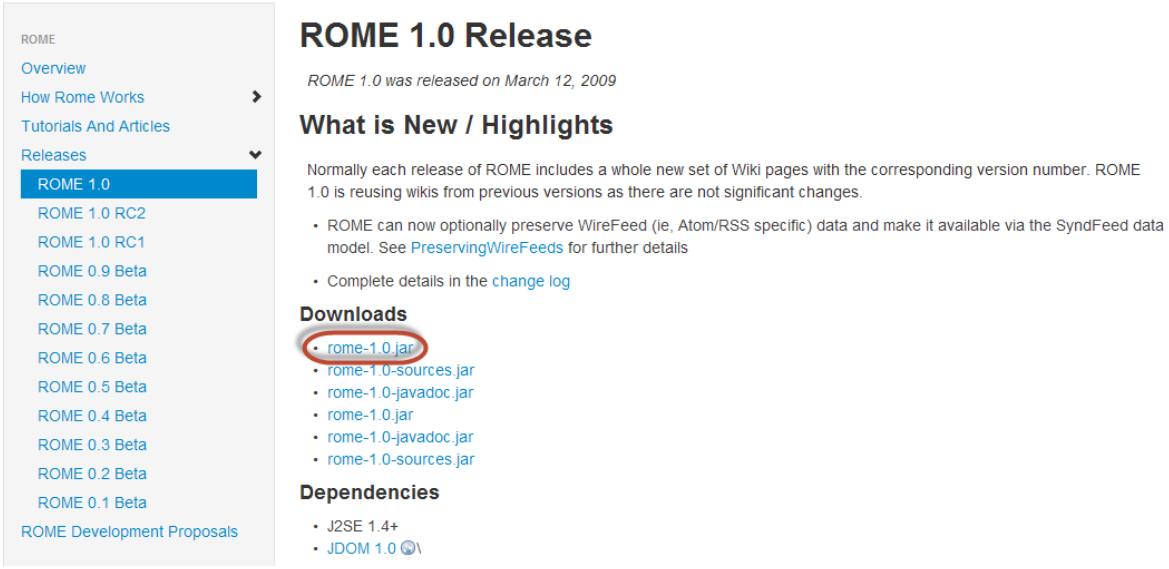



Figura 21. Descarga de ROME

### 3.2.2 Desarrollo de las clases y el módulo

Ahora explicaremos las clases más importantes que hemos desarrollado para nuestra aplicación. El primer grupo de clases son las clases que forman el modelo de la base de datos, hay que crear una clase por cada entidad de la base de datos. La siguiente clase corresponde con la clase leerRSS.java, que es la clase que se encarga de leer los distintos RSS de las páginas.

**Clases Modelo base de datos:**

Con ayuda de Hibernate Tools, creamos la conexión a la base de datos:

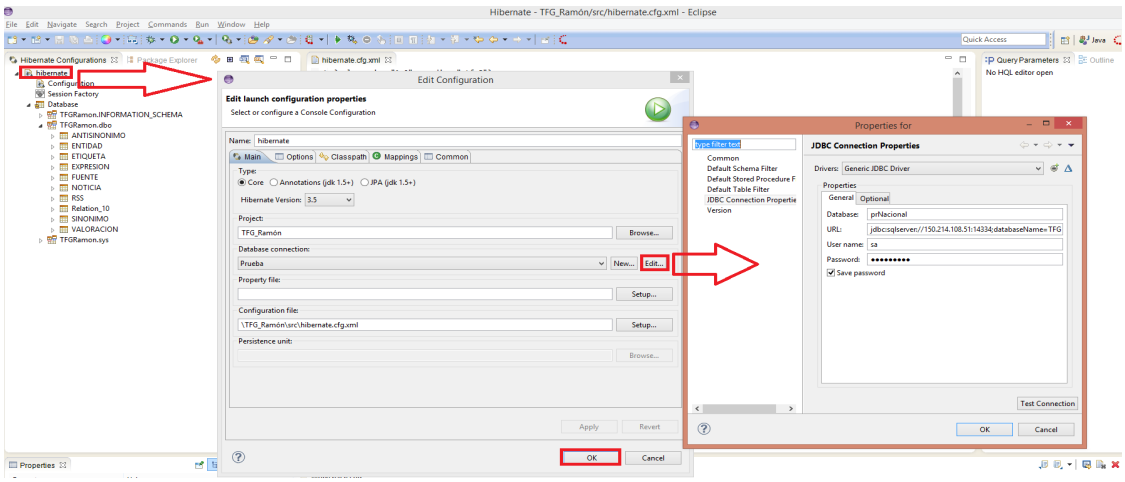



Figura 22. Configuración Hibernate

La forma de configurar hibernate es usando el fichero XML de configuración llamado hibernate.cfg.xml. Este fichero deberemos guardarlo en el paquete raíz de nuestras clases Java, en este fichero contiene distintos datos como la conexión de la base de datos y los archivos usados para el mapeo de entidades (en el siguiente párrafo indicamos como se generan).

hibernate.cfg.xml

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE hibernate-configuration PUBLIC
"-//Hibernate/Hibernate Configuration DTD 3.0//EN"
"http://hibernate.sourceforge.net/hibernate-configuration-3.0.dtd">
<hibernate-configuration>
    <session-factory>
        <!-- Database connection settings -->
        <property name="connection.driver_class">com.microsoft.sqlserver.jdbc.SQLServerDriver</property>
        <property name="connection.url">jdbc:sqlserver://150.214.108.51:14334;databaseName=TFGRamon</property>
        <!-- <property name="connection.url">jdbc:sqlserver://150.214.108.104\SQLREP;databaseName=prNacional</property> -->
        <property name="connection.username">[redacted]</property>
        <property name="connection.password">[redacted]</property>
        <!-- JDBC connection pool (use the built-in) -->
        <property name="connection.pool_size">10</property>
        <property name="hibernate.jdbc.batch_size">30</property>

        <!-- SQL dialect -->
        <property name="dialect">org.hibernate.dialect.SQLServer2008Dialect</property>

        <!-- Enable Hibernate's automatic session context management -->
        <property name="current_session_context_class">thread</property>



        <!-- Echo all executed SQL to stdout -->
        <!-- <property name="show_sql">true</property> -->
        <property name="format_sql">true</property>

        <!-- Drop and re-create the database schema on startup -->
        <!-- <property name="hbm2ddl.auto">update</property> -->

        <mapping class="modelo.Etiqueta"/>
        <mapping class="modelo.Rss"/>
        <mapping class="modelo.Antisinonimo"/>
        <mapping class="modelo.Entidad"/>
        <mapping class="modelo.Expresion"/>
        <mapping class="modelo.Fuente"/>
        <mapping class="modelo.Noticia"/>
        <mapping class="modelo.Sinonimo"/>
        <mapping class="modelo.Valoracion"/>

        <mapping resource="modelo/Etiqueta.hbm.xml" />
        <mapping resource="modelo/Rss.hbm.xml" />
        <mapping resource="modelo/Noticia.hbm.xml" />
        <mapping resource="modelo/Entidad.hbm.xml" />
        <mapping resource="modelo/Sinonimo.hbm.xml" />
        <mapping resource="modelo/Fuente.hbm.xml" />
        <mapping resource="modelo/Antisinonimo.hbm.xml" />
        <mapping resource="modelo/Expresion.hbm.xml" />
        <mapping resource="modelo/Valoracion.hbm.xml" />
    </session-factory>
</hibernate-configuration>
```

Configuración de la conexión con la BBDD

Mapeo entidades BBDD

Figura 23. Hibernate.cfg.xml

El último paso es generar los archivos .java y hbm.xml para realizar el mapeo de las tablas de base de datos a clases de java.

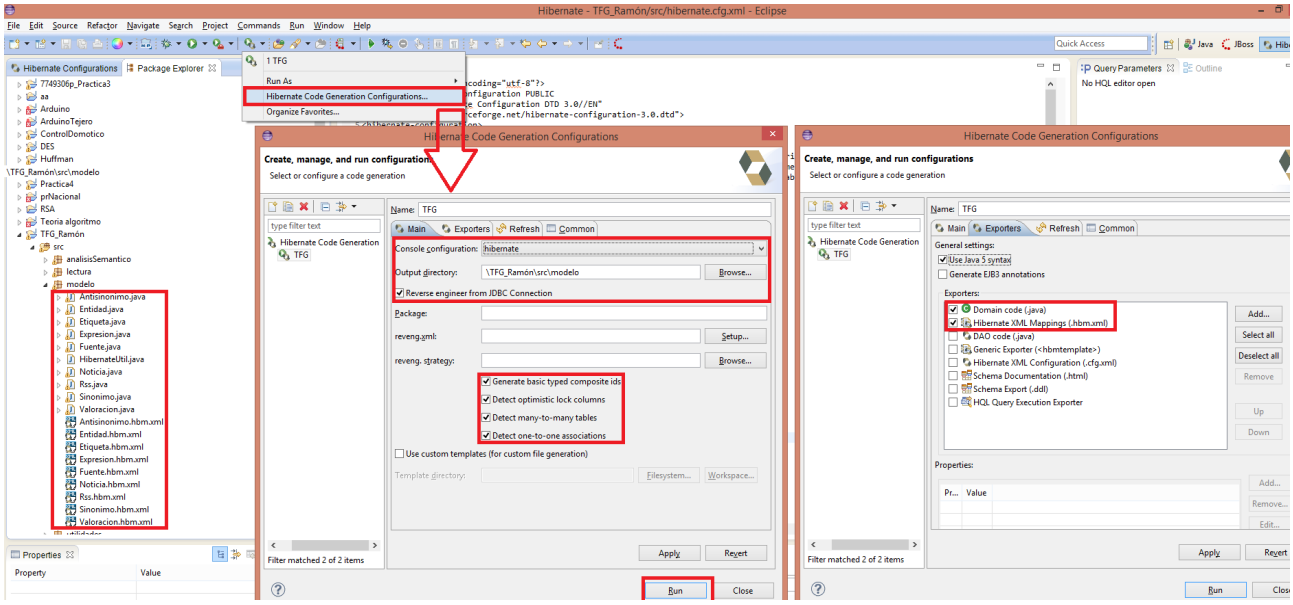



Figura 24. Generación automática de ficheros de mapeo Hibernate

**LeerRSS**
Esta es la clase que contiene los métodos necesarios para leer de los distintos RSS y guardarlos en la base de datos. Esta es la superclase del grupo de las clases fuentes.

**Métodos de la clase**
Set<Noticia> leerRSS (String RSS, String procede, Set<Noticia> lista_noticias): Este método dada la URL del RSS, el periódico al que pertenece y una lista de noticias, te devuelve la misma lista más añadiendole las noticias que se han leido del RSS.

Primero nos traemos la fuente de la base de datos, luego leemos la URL y obtenemos el código de la página. Con ayuda de la librería Rome le pasamos el XML del RSS y nos devuelve un listado de las noticias que ha leido. De ahí seleccionamos las noticias que hablan de algunas de las empresas que tenemos almacenadas y la guardamos dentro del listado final.

guardar_Noticias(Set<Noticia>lista_noticias): Guarda las noticias del listado en la base de datos.

**Clase Periodicos_Espanya:**
Este clases hereda de LeerRSS, recorre todos los RSS y los guarda en la base de datos.

**Atributos de la clase**
Set<noticia> lista_noticias: Donde se guarda la lista de noticias que se van a guardar en la base de datos.

**Métodos de la clase**
Al heredar los métodos de la clase LeerRSS conserva los métodos de esta clase.

public static void main (String [] args): Se encarga de llama al constructor de la clase.

Public constructor(): Se encarga de llamar al método leerRSS() para guardar en el listado las noticias y al final llama al método guardar_noticias() para guardarlos en la base de datos.

3.3 Tratamiento de la información

Este módulo se va a encargar de usar la información obtenido a través de el módulo de captura para obtener el RC de las empresas. Para llegar a eso hay una serie de pasos a seguir.

Lo primero es desarrollar un clasificador semántico automático que se encargue de analizar el sentimiento de la información dada y obtener su tono. En el estado de arte explicamos la dificultad que hay a la hora de construir un clasificador semántico al estar trabajando con el lenguaje Natural.

Aquí vamos a explicar el método que hemos desarrollado nosotros para la evaluación de textos al que hemos bautizado como 'Clasificador basado en expresiones', el cual explicaremos en los siguientes apartados.

El siguiente paso es obtener la Reputación Corporativa de cada empresa a partir de toda la información obtenida. El cómo obtenemos ese RC también será explicado en

siguientes apartados.

### 3.3.1 Herramientas de desarrollo

#### 3.3.1.1 Eclipse

Igual que en el anterior módulo, usaremos el lenguaje de programación JAVA para conectarnos a la base de datos y obtener la información que luego será evaluada por nuestro clasificador semántico. Este programa debe ejecutarse automáticamente cada cierto tiempo analizando el sentimiento de la información que es recogida en cada proceso de lectura asegurándose de no volver a clasificar textos que fueron analizados anteriormente

### 3.3.2 Desarrollo de las clases y el módulo

Haremos una breve explicación de las clases creadas para poner en marcha el clasificador semántico

**Clases Modelo base de datos:**

Al igual que en el anterior módulo necesitamos mapear las entidades de las bases de datos que nos hagan falta para llevar a cabo la valoración de la información que tenemos guardada.

La primera entidad es VALORACIÓN, donde guardaremos el sentimiento de la valoración una vez que la hayamos clasificado con nuestro clasificador semántico.

También es necesario mapear las entidades EXPRESIÓN y ETIQUETA, ya que son el cuerpo principal de nuestro clasificador y las necesitaremos a la hora de realizar una valoración.

Las entidades NOTICIA, ENTIDAD y SINÓNIMO que ya se usaron en el primer módulo también nos hará falta en la creación de este.

Para tener una pequeña noción de como mapear una entidad de base de datos a una clase en JAVA recomiendo mirar el apartado 3.2.2 y cualquier manual de Hibernate.

**Clase Preprocesado:**

En esta clase realizamos un tratamiento previo a la información antes de ser valorada, este tratamiento puede hacerse de distintas maneras, lo explicaremos uno a uno en los procedimientos de la clase.

**Atributos de clase**

Final string SEP_PUNTUACION: Constante que forma un string formado por distintos signos de puntuación usado para separar las frases en trozos. Los signos usados son: '?', '.' , '|' ,' /' , '!' y '\n'.

Final string LINKS: Contiene una expresión regular que reconoce si un texto es o no una URL.

Pattern patronConjunciones: Expresión regular que contiene algunas conjunciones. Ejemplo: 'y', 'pero', 'aunque', etc...

Pattern patronEnlaces: Expresión para encontrar una URL dentro de un texto, esta expresión regular, haciendo uso de la constante LINKS.

**Métodos de la clase**

Public Preprocesado: Constructor de la clase, se encarga de inicializar las variables patronEnlaces y patronConjunciones.

Public String quitarTerminosNoRelevantes(String texto): Se encarga de eliminar las palabras del texto para facilitar el análisis del texto. Por ejemplo las preposiciones, articulos y palabras mal escritas. El método devuelve el texto después de eliminar los términos relevantes.

Protected List<String> dividirEnFrases(String texto): Dado un texto lo dividimos en distintas partes para facilitar la clasificación del texto. Primero divide el texto por los signos de puntuación usando la constante SEP_PUNTUACIÓN, luego a cada fragmento de texto le aplicamos un filtro para eliminar las frases interrogativas. Y por último volvemos a separar las frases pero esta vez por conjunciones. EL método devuelve una lista con el texto una vez dividido.

Protected List<String> dividirPorPuntuacion(String texto): Dado un texto lo dividimos usando los signos de puntuación. El método devuelve una lista con el texto dividido.

Protected List<String> dividirPorConjuncion(List<String> texto): Dado una lista de trozos de un texto, lo volvemos a dividir si es posible por conjunciones. El método devuelve una lista con el texto dividido pro conjunciones.

Public String quitarNums(String texto): Dado un texto, devuelve ese mismo texto quitando los números sueltos que se encuentren en él.

Public String quitarLinks(String texto): Dado un texto, devuelve ese mismo texto eliminando las URL que se encuentren en él.

Public String quitarLetrasRepetidas(String texto): Devuelve el texto una vez eliminamos las repeticiones de letras iguales que pueden haber en el texto por algún error y . No eliminamos letras como la 'r', 'c', 'l', etc., que si pueden aparecer seguidas.

Public String quitarTildes(String texto): Remplaza las letras acentuadas en el texto por la misma letra pero sin tilde. EL método devuelve el texto sin acentuar.

**Clase cargarEtiqueta**

En esta clase guardamos el contenido de las tablas Expresión y Etiqueta.

**Atributos de la clase**

List<Etiqueta> listaEtiquetas: Contiene todas las etiquetas guardadas en la base de datos, una etiqueta contiene una clase y todas las palabras que pertenecen a esa clase.

List<Expresion> listaExpresiones: Contiene todas las expresiones de la base de datos.

Map<String,String> etiquetas: Relaciona la clase de la etiqueta con la expresión regular asociada a esa clase. Esta expresión regular indica las palabras que pertenecen a esta clase.

Map<Expresion,String> expresiones: Relaciona cada expresión con la expresión regular que lo identifica.

**Métodos de la clase**

public CargarEtiqueta(): Contructor de la clase, se encarga de llamar a los métodos cargarEtiquetas() y cargarExpresiones().

public Map<String, String> getEtiquetas(): Devuelve el contenido de la variable 'etiquetas'.

public void setEtiquetas(Map<String, String> etiquetas): Sustituye el contenido de la variable 'etiquetas' por el contenido de la variable pasado por parámetro.

public Map<Expresion, String> getExpresiones(): Devuelve el contenido de la variable 'expresiones'.

public void setExpresiones(Map<Expresion, String> expresiones): Sustituye el contenido de la variable 'expresiones' por el contenido de la variable pasado por parámetro.

public Map<String,String> cargarEtiquetas(): Se encarga de obtener el contenido de la entidad 'Etiqueta' de la base de datos y guardarlo en la variable 'listaEtiquetas'. Luego hace uso de esa variable para rellenar el contenido de 'etiquetas'.

public Map<Expresion,String> cargarExpresiones(): Se encarga de obtener el contenido de la entidad 'Expresión' de la base de datos y guardarlo en la variable 'listaExpresiones'. Luego hace uso de esa variable para rellenar el contenido de 'expresiones', las etiquetas que hay en las expresiones son sustituidas por su descripción.

**Clase MainMetodoExpresiones**

Es la clase principal que se ejecuta a la hora de valorar la información, es un programa automático que se ejecuta cada cierto tiempo y cuando haya también nueva información que clasificar.

**Atributos de la clase**

Final static int CANTIDAD_NOTICIA: Número de noticias mínimo que debe haber en la base de datos que hay sin clasificar para que nuestro sistema empiece a clasificarlos

**Métodos de la clase**

private static long cantidadInformacion(): Devuelve el número de información que hay sin valorar en la base de datos.

Public static void main(String args): Método principal del programa,se encarga de llamar al procedimiento metodoExpresiones(CargaEtiqueta ce) hasta que el número de información sin valorar sea menor que CANTIDAD_NOTICIA. Si no hay información que valorar o el número es demasiado bajo entonces el programa se duerme duarante un periodo de seis horas antes de volver a evaluar de nuevo

Private static booblean metodoExpresiones(CargaEtiqueta ce): Realiza una consulta a la base de datos para obtener un listado de tamaño máximo CANTIDAD_NOTICIA y la valora usando el método valorarInformación(list<Noticia> noticias, CargarEtique-

ta ce) de la clase MetodoExpresiones. Si el tamaño del listado es igual a CANTIDAD_NOTICIA, el método devuelve true y para indicar que sigue habiendo suficiente información que valorar, en otro caso devuelve false que significa que el programa debe esperar a que entre nueva información para volver a empezar a valorar.

**Clase MetodoExpresiones**

**Atributos de la clase**

Final string NEGACIONES: Una expresión regular con algunas palabras con un tono negativo .

**Métodos de la clase**

Public void valorarInformaciónNoticia(list<Noticia> noticias, Map<Expresion, String> expresiones): Se encarga de valorar cada noticia de las pasada por parámetro usando el método valorarCasoBaseA(...) y guardarlas en la base de datos.

Private boolean ValorarCasoBaseA( Noticia n, Map<String,NodoEntidadValoracion> terminosEncontrados, Map<Expresion, String> expresiones, String textoNoticia, Map<Expresion, String> expresionesEncontradas, List<Sinonimo> sinonimosDevolver, int flagSinonimosEncontrados): Cuenta el número de expresiones, empresas y negaciones que aparecen en el texto y con el método casoBase(...) comprueban si el texto es un caso fácil de valorar, este método devuelve true si el texto ha sido posible de valorar (es un caso base) o false en otro caso. En la variable terminosEncontrados tenemos asociado Cada empresa con la valoración obtenida al analizar el texto.

Private int ocurrencias( String texto, String patrón): Indica cuantas veces la expresión regular 'patrón' se encuentra dentro del texto 'texto'.

Private int ocurrenciasSinonimos(String texto, Map<String,NodoEntidadValoracion> map,String patron): Siendo 'patron' una expresión regular con todos los sinónimos de las empresas. El método devuelve el número de veces que se cumple la expresión regular 'patron' en 'texto', también devuelve como parámetro la variable 'map' que asocia cada empresa con su valoración inicializada a cero ya que todavia no se han valorado las expresiones encontradas.

private boolean casoBase(int ocurrenciasSinonimos,int ocurrenciasNegaciones, int contadorExpresiones): Dado el número de empresas encontradas en el texto, negaciones y expresiones dice si es o no un caso base. Decimos que un texto es un caso base en los siguientes casos.

- OcurrenciasNegaciones == 0
  - (ocurrenciasSinonimos==0 && contadorExpresiones==0)
  - (ocurrenciasSinonimos==1 && contadorExpresiones==1)
  - (ocurrenciasSinonimos>1 && contadorExpresiones==1)
  - (ocurrenciasSinonimos==1 && contadorExpresiones>1)

**Clase NodoEntidadValoracion**

Asocia cada Empresa con la valoración asociada, está clase está asociada a una

información aunque en la clase no está contemplado.

**Atributos de la clase**

Session sesion: Interfaz entre Java e Hibernate para comunicarse con la base de datos.

Entidad entidad: Empresa a la que se le asocia la información.

Int valoración: Puntuación asociada a la empresa.

Int cantidad: Nº de sumandos usados para valor.

**Métodos de la clase**

public NodoEntidadValoración(Session sesion, Entidad entidad): Constructor de la clase.

public void contarOAnyadir (int valor): Suma el valor pasado por parámetro al atributo de la clase ‘valor’ y se lo asigna a este. Suma uno a la variable 'cantidad'

public void valorar(Noticia inf): Inserta en la base de datos la valoración de la información pasada por parámetro usando la entidad y la valoración guardadas en los atributos de la clase.

3.4 Visualización de la información

3.4.1 Herramientas de desarrollo

3.4.1.1 Netbeans

NetBeans es un entorno de desarrollo integrado libre, hecho principalmente para el lenguaje de programación Java. Existe además un número importante de módulos para extenderlo. NetBeans IDE es un producto libre y gratuito sin restricciones de uso.

Netbeans IDE permite a los desarrolladores crear con rapidez aplicaciones web, empresariales, de escritorio y móviles utilizando la plataforma Java, así como JavaFX, PHP, JavaScript y Ajax, Ruby y Ruby on Rails, Groovy and Grails y C/C++.

Se ha optado por usar Netbeans IDE 8.0 para desarrollar la interfaz del sistema.

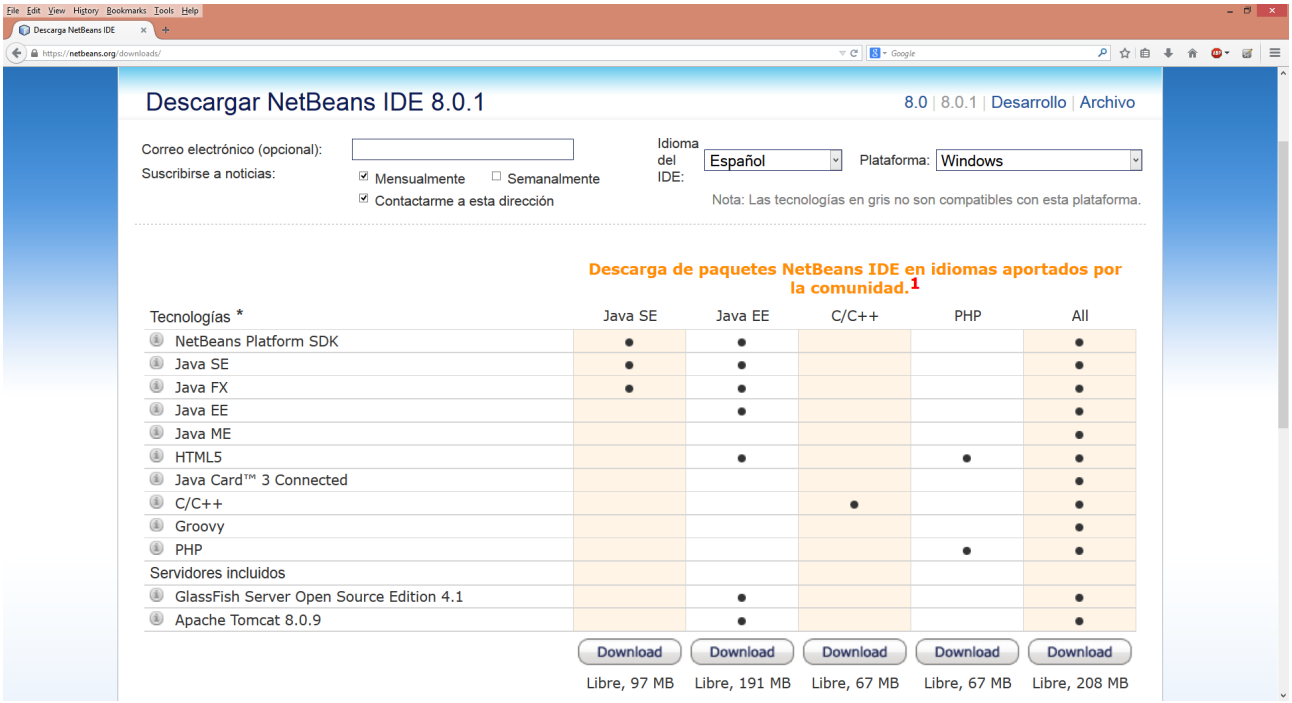



Figura 25. Página de descarga de NetBeans IDE 8.0

### 3.4.1.2 Glassfish

GlassFish es un servidor de aplicaciones de software libre desarrollado por Sun microsystems, compañía adquirida por Oracle Corporation, que implementa las tecnologías definidas en la plataforma Java EE y permite ejecutar aplicaciones que siguen esta especificación. Es gratuito, de código libre y se distribuye bajo un licenciamiento dual a través de la licencia CDDL y la GNU GPL. La versión comercial es denominada Oracle GlassFish Enterprise Server (antes Sun GlassFish Enterprise Server).

GlassFish está basado en el código fuente donado por Sun y Oracle Corporation; este último proporcionó el módulo de persistencia TopLink. GlassFish tiene como base al servidor *Sun Java System Application Server* de Oracle Corporation, un derivado de Apache Tomcat, y que usa un componente adicional llamado Grizzly que usa Java NIO para escalabilidad y velocidad.

Glassfish es el servidor de aplicaciones que usaremos para conectarnos a nuestra BBDD, viene de serie cuando se instala NetBeans IDE 8.0

### 3.4.1.3 JfreeChart

JFreeChart es un marco de software open source para el lenguaje de programación Java, el cual permite la creación de gráficos complejos de forma simple.

Esta libería es la que se ha usado para crear los gráficos que aparecen en la página web.

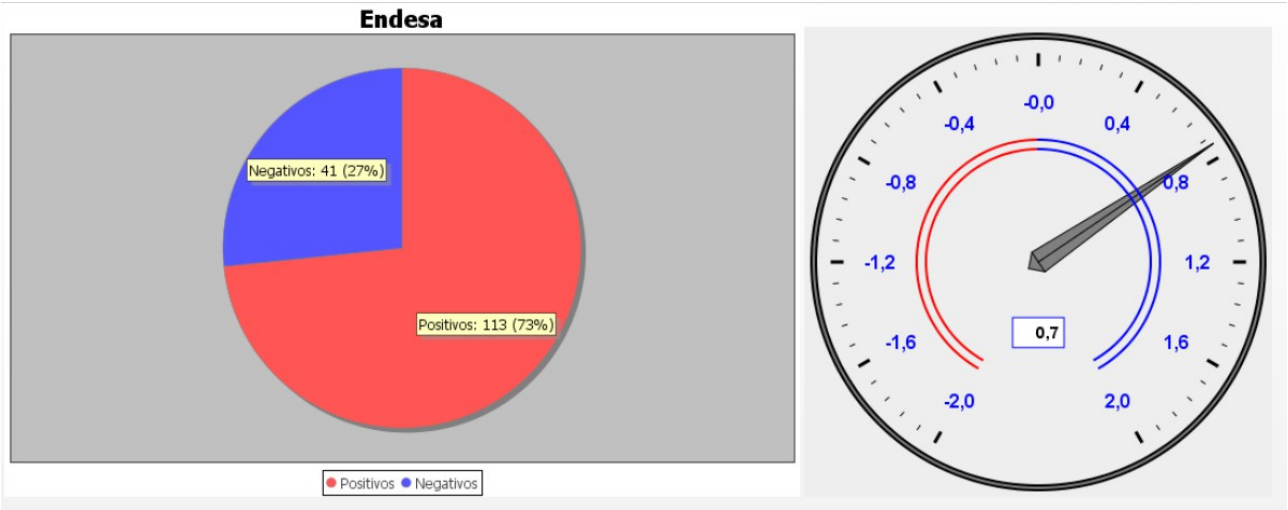



Figura 26. Gráficos interfaz

# 4.Conclusiones

## Conclusiones del Trabajo Desarrollado

En este trabajo se ha desarrollado un sistema de para determinar la reputación de las empresas en medios de comunicación online. Para ello se ha desarrollado un sistema de lectura de medios online, que permite localizar y extraer la información de los medios de comunicación online; un clasificador semántico que analizar la información recogida y la clasifica en diferentes temáticas extrayendo el sentimiento de los textos; y finalmente, una interfaz para interactuar con el usuario.

Se ha desarrollado un sistema de lectura de información en medios de información online, en concreto el sistema lee información de diarios y blog de información general, a través del sistema RSS que incorporan dichos medios. De manera que se extraiga la información de los titulares y cabecera de las noticias. Una vez que la información se ha extraído el sistema pasa dicha información al sistema de clasificación semántica.

Se ha desarrollado un clasificador semántico que analiza el sentimiento de la información y la clasifica como información positiva, neutra o negativa. El sistema está basado en el método de expresiones y tiene un porcentaje de acierto de tonalidad del 85%.

Finalmente, se ha desarrollado una interfaz para el sistema que permite gestionar los procesos anteriormente descritos de manera amigable para el usuario.

## Conclusiones del Proceso de Aprendizaje Personal

A nivel personal he encontrado interesante las siguientes actividades:

- Poner en práctica el conocimiento adquirido en la carrera a la hora de desarrollar este trabajo: Hemos tenido la oportunidad de usarlos conocimiento adquiridos a través de distintas asignaturas de la carrera, por ejemplo: Planificación de un proyecto, aprendizaje computacional, base de datos, programación, etc); y ponerlos a trabajar en grupo.

- Estudiar un nuevo campo como es la reputación corporativa: Cada vez va cobrando más importancia y eso se puede ver en cómo muchas consultoras y agencias han creado departamentos para la gestión de la reputación y la monitorización online.

- Ser capaces de desarrollar e implementar un clasificador semántico que analice el sentimiento de la información obtenida para luego poder calcular la reputación corporativa de una empresa.

- Desarrollar sistema de lectura para medios comunicación online.

## 5. Bibliografía

Una de las fuentes principales de bibliografía es la web de Hibernate. De donde he sacado toda la información para aprender a manejar la librería (http://hibernate.org/orm/documentation/).

También he usado la bibliografía de Java para despejar dudas referentes al código (http://download.oracle.com/javase/6/docs/api/overview-summary.html).

Bibliografías que me han sido útiles a la hora de desarrollar el TFG:

[1] - Ana María Casado, José Ignacio Peláez (2014). Intangible Management Monitors and Tools: Trends in the International Companies.

[2] - Leonard J Ponzi, Charles J Fombrun and Naomi A Gardberg (2011). RepTrak™ Pulse: Conceptualizing and Validating a Short-Form Measure of Corporate Reputation.

[3] - Ana Maria Casado, José Ignacio Peláez and Juan Cardona. Managing Corporate Reputation: a perspective on the Spanish market.

[4] - Alicia Vaquero Collado (2011). La reputación online en el marco de la comunicación corporativa. Una visión sobre la investigación de tendencias y perspectivas profesionales

[5] - Dolores Alonso y Victoria Pino. Reputación corporativa https://www.iae.edu.ar/antiguos/Documents/Revista21/iae21_65a66.pdf

[6] - Indurkhya N., Damerau F.J. (eds.) Handbook of natural language processing (2ed., CRC, 2010)(ISBN 9781420085921)

**Anexo I. Instrucciones DDL generar tablas en SQL Server 2012.**

A continuación el código necesario para generar las tablas en SQL Server 2012, la explicación de las tablas puede ser encontrada en el apartado 3.1.1 del TFG.

```
CREATE
  TABLE ANTISINONIMO
  (
    Antisinonimo_str VARCHAR (100) NOT NULL ,
    Sinonimo_id      SMALLINT ,
    CONSTRAINT Antisinonimo_PK PRIMARY KEY CLUSTERED (Antisinonimo_str)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO

CREATE
  TABLE ENTIDAD
  (
    Entidad_id    INTEGER NOT NULL ,
    Identidicador VARCHAR (900) NOT NULL ,
    CONSTRAINT entidad_PK PRIMARY KEY CLUSTERED (Entidad_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO

CREATE
  TABLE ETIQUETA
  (
    Etiqueta_str NVARCHAR (300) NOT NULL ,
    Definición NVARCHAR (MAX) NOT NULL ,
    CONSTRAINT Etiqueta_PK PRIMARY KEY CLUSTERED (Etiqueta_str)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO
```

```
CREATE
  TABLE EXPRESION
  (
    Expresión_id SMALLINT NOT NULL ,
    Definición   VARCHAR (MAX) NOT NULL ,
    Tono         SMALLINT ,
    CONSTRAINT Expresion_PK PRIMARY KEY CLUSTERED (Expresión_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO

CREATE
  TABLE FUENTE
  (
    Fuente_id BIGINT NOT NULL ,
    Descripción VARCHAR (1000) NOT NULL ,
    CONSTRAINT Fuente_PK PRIMARY KEY CLUSTERED (Fuente_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO

CREATE
  TABLE NOTICIA
  (
    Informacion_id BIGINT NOT NULL ,
    Fecha DATE NOT NULL ,
    Hora TIME ,
    Titulo VARCHAR (1000) NOT NULL ,
    Cabecera TEXT ,
    URL VARCHAR (300) NOT NULL ,
    Fuente_id BIGINT NOT NULL ,
    CONSTRAINT noticia_PK PRIMARY KEY CLUSTERED (Informacion_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
```

```
  ON "default"
GO

CREATE
  TABLE RSS
  (
    URL VARCHAR (500) NOT NULL ,
    Fuente_id BIGINT NOT NULL ,
    CONSTRAINT RSS_PK PRIMARY KEY CLUSTERED (URL)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO

CREATE
  TABLE Relation_10
  (
    SINONIMO_Sinonimo_id SMALLINT NOT NULL ,
    NOTICIA_Informacion_id BIGINT NOT NULL ,
    CONSTRAINT Relation_10_PK PRIMARY KEY CLUSTERED
(SINONIMO_Sinonimo_id,
    NOTICIA_Informacion_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO

CREATE
  TABLE SINONIMO
  (
    Sinonimo_id  SMALLINT NOT NULL ,
    Sinonimo_str VARCHAR (100) NOT NULL ,
    Valoración   SMALLINT ,
    Entidad_id   INTEGER NOT NULL ,
    CONSTRAINT SINONIMO_PK PRIMARY KEY CLUSTERED (Sinonimo_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
```

```
  ON "default"
GO

CREATE
  TABLE VALORACION
  (
    valoración_id BIGINT NOT NULL ,
    valor FLOAT ,
    Entidad_id INTEGER NOT NULL ,
    Informacion_id BIGINT NOT NULL ,
    modo SMALLINT ,
    CONSTRAINT valoracion__PK PRIMARY KEY CLUSTERED (valoración_id)
WITH
  (
    ALLOW_PAGE_LOCKS = ON ,
    ALLOW_ROW_LOCKS  = ON
  )
  ON "default"
  )
  ON "default"
GO
CREATE UNIQUE NONCLUSTERED INDEX
VALORACION__IDX ON VALORACION
(
  Informacion_id
)
ON "default"
GO

ALTER TABLE Relation_10
ADD CONSTRAINT FK_ASS_1 FOREIGN KEY
(
SINONIMO_Sinonimo_id
)
REFERENCES SINONIMO
(
Sinonimo_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO

ALTER TABLE Relation_10
ADD CONSTRAINT FK_ASS_2 FOREIGN KEY
(
NOTICIA_Informacion_id
)
REFERENCES NOTICIA
(
Informacion_id
```

```
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO

ALTER TABLE VALORACION
ADD CONSTRAINT Relation_13 FOREIGN KEY
(
Informacion_id
)
REFERENCES NOTICIA
(
Informacion_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO

ALTER TABLE SINONIMO
ADD CONSTRAINT Relation_14 FOREIGN KEY
(
Entidad_id
)
REFERENCES ENTIDAD
(
Entidad_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO

ALTER TABLE NOTICIA
ADD CONSTRAINT Relation_15 FOREIGN KEY
(
Fuente_id
)
REFERENCES FUENTE
(
Fuente_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO
```

```
ALTER TABLE RSS
ADD CONSTRAINT Relation_16 FOREIGN KEY
(
Fuente_id
)
REFERENCES FUENTE
(
Fuente_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO

ALTER TABLE ANTISINONIMO
ADD CONSTRAINT Relation_2 FOREIGN KEY
(
Sinonimo_id
)
REFERENCES SINONIMO
(
Sinonimo_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO

ALTER TABLE VALORACION
ADD CONSTRAINT Relation_5 FOREIGN KEY
(
Entidad_id
)
REFERENCES ENTIDAD
(
Entidad_id
)
ON
DELETE
  NO ACTION ON
UPDATE NO ACTION
GO
```